Biased Behavior in Web Activities

From Understanding to Unbiased Visual Exploration

Eduardo Graells-Garrido

TESI DOCTORAL UPF / ANY 2015

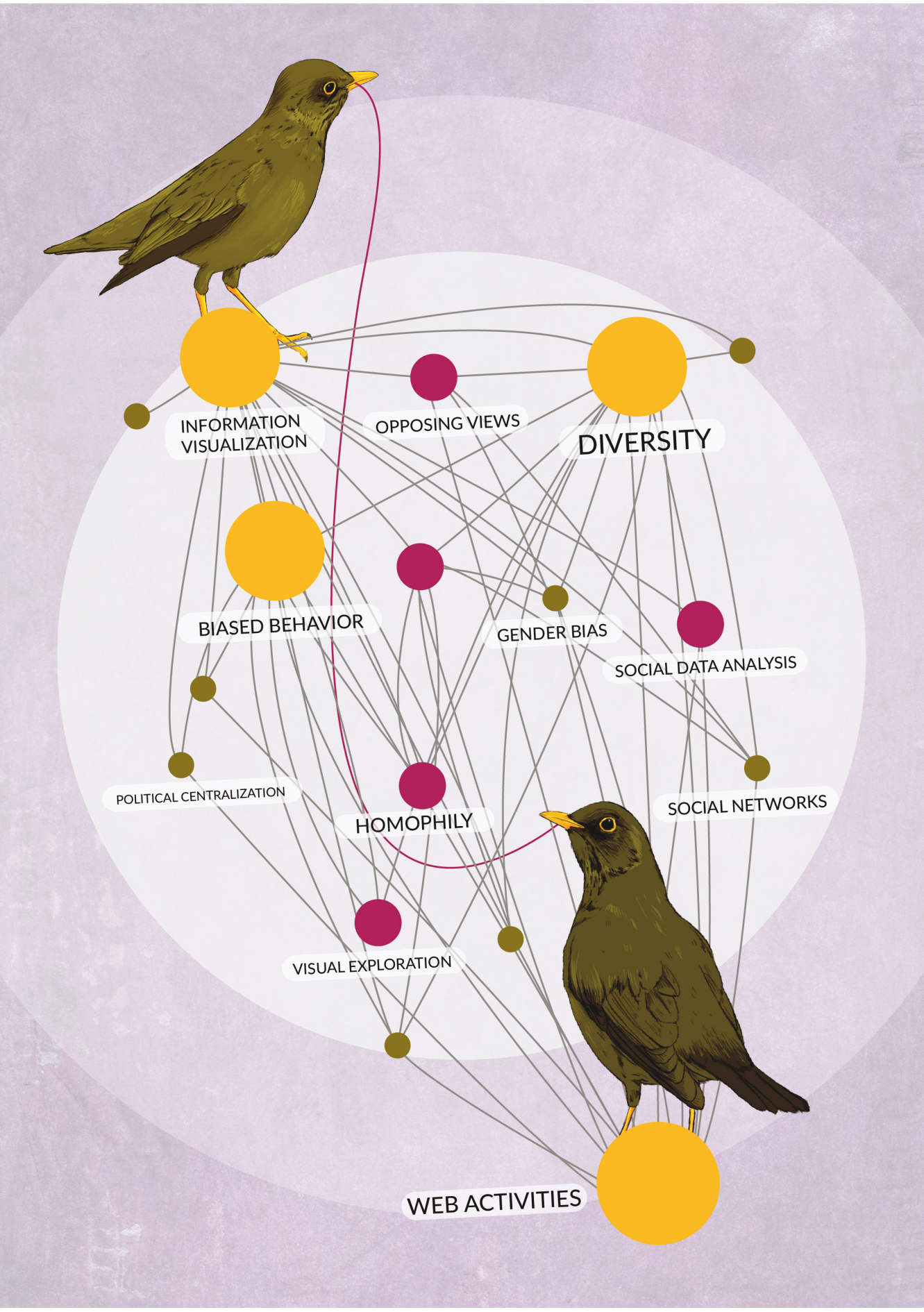
DIRECTORES DE LA TESI
Prof. Dr. Ricardo Baeza-Yates, Universitat Pompeu Fabra
Dr. Mounia Lalmas, Yahoo Labs

What transforms this world is – knowledge. Do you see what I mean? Nothing else can change anything in this world. Knowledge alone is capable of transforming the world, while at the same time leaving it exactly as it is. When you look at the world with knowledge, you realize that things are unchangeable and at the same time are constantly being transformed. You may ask what good it does us. Let's put it this way – human beings possess the weapon of knowledge in order to make life bearable. For animals such things aren't necessary. Animals don't need knowledge or anything of the sort to make life bearable. But human beings do need something, and with knowledge they can make the very intolerableness of life a weapon, though at the same time that intolerableness is not reduced in the slightest. That's all there is to it.

— Yukio Mishima

Dedicated to the loving memory of Eduardo Graells-Salazar.

1956 – 2014


INFORMATION VISUALIZATION
OPPOSING VIEWS
DIVERSITY
BIASED BEHAVIOR
GENDER BIAS
SOCIAL DATA ANALYSIS
POLITICAL CENTRALIZATION
HOMOPHILY
SOCIAL NETWORKS
VISUAL EXPLORATION
WEB ACTIVITIES

No solo no hubiéramos sido nada sin ustedes, sino con toda la gente que estuvo a nuestro alrededor desde el comienzo; algunos siguen hasta hoy. Gracias totales.
– Gustavo Cerati

ACKNOWLEDGMENTS

Pajarito, whose unconditional love, support, and inspirational ideas, helped me to become what I am now. I love you.

My family for having sacrificed so much. An ocean was not enough to diminish their love. Father, I'm still learning from you. I miss you.

My advisors, Ricardo Baeza-Yates and Mounia Lalmas. They have been role models of wisdom, integrity, humbleness, and guidance. I wish one day to be like them and support others in the same way.

My collaborators: Alejandro Jaimes (*Linspiration*), Bárbara Poblete (*#Santiago is not #Chile*), Daniele Quercia (*Data Portraits*), and Filippo Menczer (*First Women, Second Sex*). Each one of you gave me something unique.

My adventure partners: Luca Chiarandini, Diego Sáez-Trumper, Luz Rello, María Arteaga, and *Jayermeister*. Our adventure has not ended, I promise.

My Chilean friends who have helped me with econometrics and regressions: Sergio "Skewness" Salgado and Denis Parra. And thanks to Andrés Lucero who always helped me to find participants for the experiments.

Yahoo Labs, Barcelona Media and Universitat Pompeu Fabra for supporting me through a scholarship, and allowing me to finish my thesis in Santiago, Chile. There is no better place to do a PhD. A special mention to my co-generational fellows: Ruth, Michele and Janette. You inspired me in many ways.

Adeyemi Ajao and Ian Brillembourgh for giving me the opportunity of working in the USA. Thanks to Raúl "Dr. Van Rainbows" Aliaga for introducing me to them, as well as visiting us in Barcelona to share one of the best *18s* ever.

Hernán Orellana, Andrés Leiva and Pablo García, who gave me the chance to defend this thesis in Barcelona while working at Telefónica I+D.

Paula Pérez for the beautiful illustration of *zorzales* with the thesis keywords.

To Barcelona and its people for giving us new colors.

ABSTRACT

Current trends in Web content point towards personalization of content, which would not be a problem in an uniform, unbiased world, but our world is neither uniform nor unbiased. In this dissertation, we hypothesize that systemic and cognitive biases that affect users in the physical world also affect them when exploring content on the Web. We propose that biased behavior can be encouraged to be reduced through a holistic process that includes bias quantification, algorithmic formulation, and user interface design. Those three parts of our proposed process are implemented used Web Mining techniques, guided by Social Science, and presented to users in Casual Information Visualization systems. In particular, we follow a transversal approach where we apply this process, with different profundity levels, in specific case studies on Wikipedia and Twitter.

As result, we observe that biases from the physical world are indeed reflected on Web platforms, and this reflection affects content, perception and behavior of users. From this observation, and through the cross-sectional analysis of the case studies, we conclude: 1) that Web Mining tools are effective to measure and detect biased behavior; 2) that Information Visualization techniques aimed at non-experts encourages unbiased exploration of content; and 3) that one size does not fit all, and that in addition to the social, behavioral, and cultural contexts, biases should be accounted for when designing systems.

RESUMEN

Las tendencias actuales en la Web apuntan hacia la personalización de contenido, lo que no sería un problema en un mundo uniforme y sin sesgos, pero nuestro mundo no es ni uniforme ni libre de sesgos. En esta tesis, planteamos la hipótesis de que los sesgos sistémicos y cognitivos que afectan a las personas en el mundo físico también afectan el comportamiento de éstas al explorar contenido en la Web. Proponemos que es posible fomentar una disminución en el comportamiento sesgado a través de una mirada holística que incluye cuantificación de sesgos, formulación de algoritmos, y diseño de interfaces de usuario. Estas tres partes del proceso propuesto son implementadas utilizando técnicas de Minería de la Web. A su vez, son guiadas por las Ciencias Sociales, y presentadas a través de sistemas Casuales de Visualización de Información. Seguimos un enfoque transversal en el cual se aplica este proceso con diferentes niveles de profundidad a lo largo de tres casos de estudio en Wikipedia y Twitter.

Como resultado, observamos que los sesgos presentes en el mundo físico efectivamente se ven reflejados en plataformas Web, afectando el contenido, la percepción y el comportamiento de las personas. A través del análisis transversal de los casos de estudio, se presentan las siguientes conclusiones: 1) las herramientas de Minería de la Web son efectivas para medir y detectar comportamiento sesgado; 2) las técnicas de Visualización de Información enfocadas en personas no expertas fomentan el comportamiento no sesgado; y 3) no existen soluciones universales, y en adición a los contextos sociales y culturales, los sesgos deben ser considerados a la hora de diseñar sistemas.

RESUM

Les tendències actuals en la Web apunten cap a la personalització de contingut, el que no seria un problema en un món uniforme i sense biaixos, però el nostre món no és ni uniforme ni lliure de biaixos. En aquesta tesi, plantegem la hipòtesi que els biaixos sistèmics i cognitius que afecten les persones també afecten el comportament d'aquestes en explorar contingut a la Web. Proposem que és possible fomentar una disminució en el comportament esbiaixat a través d'una mirada holística que inclou quantificació de biaixos, formulació d'algorismes, i disseny d'interfícies d'usuari. Aquestes tres parts del procés proposat són implementades utilitzant tècniques de Mineria de la Web. Al seu torn, són guiades per conceptes de Ciències Socials, i presentades a través de sistemes Casuals de Visualització d'Informació. Seguim un enfocament transversal en el qual s'aplica aquest procés amb diferents nivells de profunditat al llarg de tres casos d'estudi en Wikipedia i Twitter.

Com a resultat, observem que els biaixos presents en el món físic efectivament es veuen reflectits en plataformes Web, afectant el contingut, la percepció i el comportament de les persones. A través de l'anàlisi transversal dels casos d'estudi, es presenten les següents conclusions: 1) les eines de Mineria Web són efectives per mesurar i detectar comportament esbiaixat; 2) les tècniques de Visualització d'Informació enfocades a persones no expertes fomenten el comportament no esbiaixat; i 3) no hi ha solucions universals, i en addició als contextos socials i culturals, els biaixos han de ser considerats a l'hora de dissenyar sistemes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

1

INTRODUCTION

The main topic of this dissertation is the study of biases in Web activities. Broadly speaking, there are two kinds of biases: cognitive and systemic. Both affect how humans behave when performing activities in physical and virtual worlds: cognitive biases deviate human judgment, making decisions to be illogical or irrational [HNA05], while systemic biases show a tendency to favor specific outcomes of institutions composed by humans, from industries to governments.

In this dissertation we explore the effects of such biases on specific Web activities. By performing case studies, we gradually move from understanding biased behavior to proposing algorithms and user interfaces that aim to encourage unbiased behavior.

1.1 MOTIVATION

Web activities, as any other human activity, are affected by cognitive and systemic biases. Whether the presence of those biases is good or bad depends on the context. For instance, some biases affect how people relate to others, and while arguably it is good to be related with a diverse set of people, sometimes being connected with like-minded people is easier because of a common cultural and social background. But being connected only with like-minded people tends to polarize views. This happens in physical and virtual worlds alike, yet the Web is vast, which formulates the question whether its vastness amplifies biases and their effects, and what are the effects of those biases in Web systems.

Without emitting judgment about good or bad biases, our motivation is to find user-centered ways to encourage unbiased behavior. Note that the word

encourage means that a change in behavior is optional, as it is a choice to behave in specific ways. What we propose is to empower users to make these choices in conscious and rational ways.

1.2 CONTEXT

Biases in behavior have been researched primarily in *Social and Behavioral Sciences*. Computer Science has tried to diminish bias effects on activities from two fronts: on one hand, *Web Mining* techniques support the analysis of biased activities, as well as the definitions of algorithms to diminish bias effects. On the other hand, *Information Visualization* (a subset of *Human-Computer Interaction*) allows to create visual representations of user generated content that could allow a unbiased exploration. These areas and their pairwise intersections are displayed in Figure 1.1. The pairwise intersections are composed of: *Visual Analytics for the Web* (see a PhD thesis by Pascual Cid [Pas10]), *Social Data Analysis* as defined by Wattenberg and Kriss [WK06]), and *Computational Social Science* (see an essay by Watts [Wat13]). This dissertation lies in the intersection of the three areas: we use Web Mining techniques to perform bias analysis, which we support by doing exploratory analysis with Visual Analytics for the Web. At the same time, we design Information Visualization user interfaces, with a design informed by a process of Social Data Analysis. Our work is contextualized into Social Science when appropriate, and we evaluate our analysis using methods from Computational Social Science.

1.3 RESEARCH QUESTION AND GOALS

Our research question is:

In biased Web activities, how can we encourage unbiased behavior?

Based on this question, we will perform case studies with the following goals: to *understand biases*, to *encourage change in behavior*, and to *explain when our results can be applied in current and future Web platforms.*

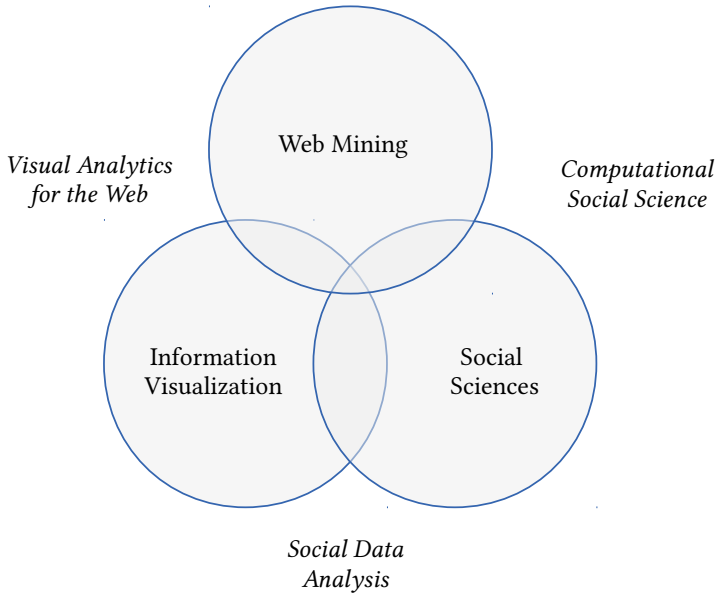


Figure 1.1: This dissertation is situated in the intersection of three research areas: Web Mining, Information Visualization, and Social Sciences.

1.3.1 *Understand Biased Behavior and its Consequences on the Web*

By understanding we mean to have practical knowledge about the expression of biases on Web activities. Cognitive and systemic biases might be well documented on the literature, and thus, our research is not aimed at discovering biases, but to understand in which contexts they influence user behavior. Moreover, we also want to understand what are the consequences of biased behavior on the Web. This is important given current trends in Web platforms aiming at *personalized user experiences*, which include recommendations and search results. For instance, the *filter bubble* phenomena described by Pariser [Par11] is about systems that hide challenging or non-agreeable information to users,

keeping them in a bubble of filtered content–with filters that were built according to each user's behavior in information seeking.

1.3.2 *A Holistic Approach to Encourage Unbiased Exploration*

Bias behavior happens when users make a choice that is deviated from the rational, or arguably best, choice from a pool of options. For instance, in the presence of challenging information, users tend to discard the information that is against their beliefs, even if it is factual and the agreeable information is wrong or false. This *selective exposure* happens because users want, unconsciously, to avoid *cognitive dissonance* [Fes62], an uncomfortable state of mind. Then, our goal is to learn if unbiased behavior can be encouraged. This research topic is not new, but previous efforts have focused on, either algorithmically providing unbiased content, or by displaying information by making the opposing information (or equivalent concept, as it depends on the bias) explicit. However, a holistic view on the Web platform that delivers the content has not been approached yet. We propose such holistic view, where first we perform a deep analysis of a specific Web activity. The analysis informs the design of, first, algorithms that provide unbiased information, and, second, visualization techniques that are used to display this information.

In particular, our designs aim to have the following traits of socially aware systems defined by Donath [Don14]:

- *Be innovative*: "Explore extraordinary possibilities".
- *Be legible*: "Bring clarity to a complex and abstract environment".
- *Be socially beneficial*: "Support the emergence of desirable social norms and cultures".

By having these traits in mind when designing systems, we believe we can encourage unbiased exploration in Web activities.

1.3.3 *Unbiased Exploration: Who, When, and How?*

Our previous goals relate to aggregated user generated content, because through aggregation we are able to learn and analyze patterns in behavior. However, no two users are equal, and the study of individual differences will prove to be use-

ful when trying to convert our understanding gained through the dissertation into guidelines for algorithm and user interface design in biased contexts. These guidelines will help to explain *who* should be approached by designs like ours, *when* to employ such algorithms and designs, and *how* to perform a case study and design systems to encourage unbiased exploration in biased contexts.

1.4 APPROACH, METHODS AND RESULTS

We follow a transversal approach where we perform case studies for different biased Web activities. In those case studies, we gradually start to think about design of user interfaces that would help users to behave in a less biased way. In total, we perform three case studies about the following biases: *gender bias*, *political centralization*, and *political homophily*.

1.4.1 *Gender Bias in Wikipedia Characterization of Historical Figures*

Our first case study, *"Gender Bias in Wikipedia"* (Chapter 3) is about gender bias on characterization of historical people in Wikipedia. Here, we target the community of Wikipedia contributors, who, to be able to contribute content to Wikipedia, follow specific guidelines of subject notability and neutrality. However, as we find, these guidelines are not sufficient to avoid bias in content. We make use of computational linguistic methods to quantify how different the characterization of women is to those of men, as well as to find qualitative explanations based on social theory that will help to define guidelines to avoid such biases.

Main Results

We perform a careful analysis where we considered that Wikipedia is an encyclopedia that reflects world knowledge, including biases that affect the physical world. Having that in mind, we find more similarities than differences in the characterization of women and men. These differences can be strongly associated to sexist behavior according to social theory. Not all of them are inherent to Wikipedia, as some of them are reflections of our already biased society. However, some differences can be associated to the community in Wikipedia. One

of them gave the name to the case study: *"first women"* is a characteristic phrase to describe women in Wikipedia biographies [GLM15].[1]

Although in this first case study we do not design a user interface, it builds the foundation of our methods to understand biased behavior in the following chapters of the dissertation. We decided to change the platform of study in the next studies, because the user population of Wikipedia is composed mainly by expert users, it is already biased towards men participation [Lam+11], and this bias in their population goes beyond exploration of content.

1.4.2 *Political Centralization in Developing Countries*

In this case study, named *Encouraging Diversity Awareness* (Chapter 4), we perform analysis of biases in micro-blogging platforms, as this allows us to understand biased behavior of a community of non-experts. The study tackles the systemic bias of *political centralization* [Kol13] in Chile, which tends to favor the capital, Santiago, in media outlets, public policy, and economical powers [GK08]. In the context of political elections in 2012 (*#municipales2012*), we crawled micro-posts published by the population, and analyzed if the centralization from the physical world is reflected on the micro-blogging platform Twitter.

In comparison to the first study, this one goes further, as we define an algorithm that, given a timeline of micro-posts, generates a summary timeline with ensured geographical diversity, focusing on informativeness and interestingness. Later, the algorithm output is evaluated with users. Using known visualization techniques applied in news contexts, we define a user interface to make the inherent diversity in the dataset visible, so "*users can see it*". To reach end-users with this design and perform an evaluation "in the wild" [Cra+13], we create a social platform in Twitter itself, through the usage of a social bot named *@todocl*. This bot is used to promote the timelines generated by our algorithm, and to share links to our visual interface.

1 Although these results are not yet published, a pre-print is available, as referenced.

Main Results

Through a characterization and analysis of user generated content [Gra14], we are able to confirm than Chilean users in Twitter behave in a politically centralized way [GL14]. Moreover, we find that centralization affects algorithmic processing of information [GP13], and that users have different perception of timelines depending on the geographical origin with respect to to political centralization. By evaluating interaction data with our prototype deployment, we are able to understand user differences with respect to engagement with the site. Confirming the results of the controlled user study of timeline perception, the condition of being from a centralized location influences user behavior. Furthermore, our proposed visualization-based interface effectively engages users and encourages them to explore content, as indicated by statistical analysis of the logged data.[2]

1.4.3 *Political Homophily in Micro-blogging Platforms*

In this case study, named *Encouraging Exploration with Data Portraits* (Chapter 5), we study the cognitive bias of homophily [MSC01], which is the tendency of individuals to create ties with like-minded people. This problem has been approached in several ways on the literature, yet results have not been satisfactory (for a extensive review on this subject, see the doctoral thesis by Munson [Mun12]). Some factors include the selective exposure mechanism and that people just do not value diversity as much as they could.

On previous attempts in the literature, the approach taken has been direct, *i. e.*, users are presented with recommendations that are known to activate selective exposure, or confronting information is displayed in a different way, but it is confronting anyway. We propose an indirect approach, where we consider partial homophily: we recommend people who have opposite political stances, but that share interests that will not activate the selective exposure mechanism (*e. g.*, music, sports, and so on). In a holistic view of the system, we incorporate the recommender system in a system that displays its recommendations in an engaging and attractive way. This way is through a *data portrait* [Don+10], a

2 These results are not yet published.

visual depiction of a user's profile, where we emphasize user interests to be linked with recommendations.

As in the second case study, we explore the Chilean community in Twitter discussing political issues in the context of presidential campaigns for the elections of 2013 (*#presidenciales2013*). In particular, we focus on one specific issue: *abortion*, for which we analyze the surrounding discussion, and evaluate if Chilean users present homophilic behavior when discussing about abortion. Based on a qualitative analysis of user feedback, we refine our algorithm by introducing *intermediary topics* [GLQ14], which formalize the idea of shared interests by using topic modeling [BNJ03]. Following the redefinition of the recommender algorithm, we create a new design for the data portrait, as well as a new depiction of recommendations. Then we incorporate this new implementation into the social platform from the previous case study, *@todocl*, and release an implementation "in the wild", for anyone to register and use.

Main Results

When characterizing population behavior, we find that users present homophilic interactions in terms of the studied political issue [GLQ13].[3] By evaluating interaction data with our deployment, we observe that individual differences in terms of political content in user profiles affect how users interact with our application, even if the application is not about politics. We also observe that behavioral differences in terms of informational behavior play a role when interacting with the data portrait and the recommendations.

In general, we find that the usage of visualization encourages exploration of recommendations of opposing people regardless of the algorithm used to generate those recommendations. Conversely, we also find that politically homophilic recommendations still are a prevalent factor when accepting recommendations. In particular, we find that our approach encourages a *conscious decision-making process* for politically involved users when facing diverse recommendations.

3 Although these results are not yet published, a preliminary report of the study is available, as referenced.

1.5 CONTRIBUTIONS

After performing the three case studies, in Chapter 6 we analyze the main contributions of this dissertation through the analysis of the common factors and design implications derived from a cross-sectional analysis of study results.

The following is a summary of the main findings:

1. *Web Mining Tools are Effective when Measuring Bias in Content and Behavior.* We find that biases from the physical world are effectively reflected on the entire Web content life cycle. This reflection is confirmed and quantified by using tools from the several disciplines belonging to Web Mining.
2. *Social Sciences Frame and Guide the Analysis.* Even though this dissertation lies primarily on Computer Science, the Social Science point of view is needed to guide analysis and design. Without a social framework, an effective *pluralist design* [Bar10] would be not feasible.
3. *Unbiased Algorithms are Necessary but not Sufficient.* Through the dissertation we develop algorithms that theoretically solve the problem of exploring biased information spaces, because they provide diverse or unbiased content. However, even when presented with unbiased information, users either do not value this or do not see the diversity provided by the algorithm.
4. *Information Visualization Encourages Exploration of Diverse Content.* The statistical analysis of user behavior from the case studies confirms that visualization-based user interfaces encourage exploratory behavior in biased scenarios.
5. *One Size does not Fit All.* We find that behavioral individual differences are important when designing systems targeted at end-users. Additionally, we find that biases introduce specific differences that must be accounted for, as biases influence the social and cultural contexts surrounding individuals.
6. *User Engagement Allows to Measure Differences in Behavior "In the Wild".* The systems we propose are not task-based and are focused on end-users. We find that user engagement metrics [LOY14] allow us to perform quantitative evaluation of user behavior, specially in our "in the wild" deployments.

The analysis we perform in this dissertation allow us to understand *who* should be targeted with proposals like ours; *when* these designs should be used; and *how* the analysis of biases should inform user interface design and implementation. We contextualize our results in social theory to explain the *why* behind them. However, we only theorize about it, and longitudinal, qualitative studies are still needed in future work as empathy towards the targeted communities and their social contexts is always necessary.

The results of the thesis have produced the following publications:

[GL14] Eduardo Graells-Garrido and Mounia Lalmas. "Balancing Diversity to Counter-measure Geographical Centralization in Microblogging Platforms *(short paper)*". In: *25th ACM Conference on Hypertext and Social Media* (2014).

[GLM15] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. "First Women, Second Sex: Gender Bias in Wikipedia". In: *arXiv preprint arXiv:1502.02341* (2015).

[GLQ13] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. "Data Portraits: Connecting People of Opposing Views". In: *arXiv preprint arXiv:1311.4658* (2013).

[GLQ14] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. "People of opposing views can share common interests". In: *Proceedings of the companion publication of the 23rd international conference on World Wide Web (poster)*. International World Wide Web Conferences Steering Committee. 2014, pp. 281–282.

[GP13] Eduardo Graells-Garrido and Bárbara Poblete. "#Santiago is not #Chile, or is it?: a model to normalize social media impact". In: *Proceedings of the 2013 Chilean Conference on Human-Computer Interaction*. ACM. 2013, pp. 110–115.

[Gra14] Eduardo Graells-Garrido. "Ornitología Virtual: Caracterizando a #Chile en Twitter". In: *Socializar Conocimientos N 2: Observando a Chile desde la Distancia*. Ed. by Lorena B Valderrama *et al.* 2014.

This list includes pre-prints that have not been published at the time of this thesis submission.

2

BACKGROUND

In this chapter we describe the core ideas, concepts and definitions that shape the context where this dissertation lies in.

2.1 PLATFORMS

Here we describe the two analyzed platforms in this dissertation.

2.1.1 *Wikipedia*

Wikipedia[1] is an open encyclopedia where anyone can contribute content.[2] In traditional encyclopedias, a staff of experts in specific areas takes care of writing, editing and validating content, while in Wikipedia a community of volunteers is responsible. Each Wikipedia article can contain links to other articles, and its content can be edited by several people. These edits are recorded into an *edit history*, creating a timeline of changes for each article, noting who made the change and what was changed. Figure 2.1 shows a screenshot of a Wikipedia article.

There is a extensive body of research built upong Wikipedia (see a survey by Okoli *et al.* [Oko+14]), covering topics like growth [AMC07], dynamics [Rat+10], accuracy of content [Gil05; Ros06], participation [Mor+13; CT14], generation of structured data [Leh+14b], analysis of historical figures [Ara+12], among others. Of particular interest for this dissertation is the *gender-gap* on Wikipedia,

1 http://en.wikipedia.org

2 Wikipedia is available for several languages. In this dissertation we study the English version.

Figure 2.1: Screenshot of a Wikipedia article (*Magellanic Penguin*).

where women represent only 16% of editors [HS13]. However, an analysis of how this gap affects Wikipedia content has not been done yet in terms of gender. A related concept is the *self-focus bias* introduced by Hecht and Gergle [HG09], which is about cultural bias on Wikipedia.

Dissertation Context

In Chapter 3 we analyze article content to compare how women are described in biographies, and see whether these descriptions differ from those of men.

2.1.2 *Twitter*

Twitter[3] is a micro-blogging platform where users publish status updates called micro-posts or *tweets*, each with a maximum length of 140 characters. Each user has a *timeline*, a list of tweets in reverse chronological order. A tweet can be marked as favorite, replied, or *retweeted*. A *retweet* is a re-publication of a tweet into the timeline of the retweeting user. Users can *follow* other users, establishing directed connections between them. When user A follows user B, tweets and *retweets* made by B will show up in A's timeline. Users can annotate tweets using *hashtags*, *i. e.*, keywords that start with the hash character `#`. To mention another user, her or his username must be prefixed with the character `@` (*e. g.*, *@A*). Figure 2.2 shows the screenshot of a Twitter profile.

Given its structure and evolution, more than a social network, Twitter is an *information network* [Kwa+10]. Several research areas have analyzed Twitter: information filtering [DCC11], recommender systems [Che+12], geolocation of users [Rou+13], cultural differences [GMQ14], crisis mapping [Mac+11], among others.

Dissertation Context

Two case studies (Chapters 4 and 5) study political discussion on Twitter. In the first one we study *political centralization*; in the second we study *homophily*.

2.2 COGNITIVE, SYSTEMIC AND GENDER BIASES

In this section we describe the biases relevant to this dissertation.

2.2.1 *Gender Bias*

The first bias we focus on is *gender bias*. In particular, we address bias on language to describe women in comparison to men in Wikipedia. Lakoff [Lak73] pioneered this area, by analyzing how language used to refer to women reflects women's inferior role in society. Analysis of language and gender adopts four

3 https://twitter.com

Figure 2.2: Screenshot of a Twitter profile (*Yahoo Labs*).

approaches to gendered speech: *deficit* (women's language is inferior to the normative men's language), *dominance* (women is seen as subordinate), *difference* (women and men have different subcultures), and *dynamic* (language is an evolving social construction which depends on many factors) [Coa04]. This bias is present, we argue, because women are not treated as equals, but as *others*, as stated by the seminal work *The Second Sex* by Simone de Beauvoir [De 12].

Dissertation Context

In Chapter 3 we analyze encyclopedic content in Wikipedia. This has been partly studied before for Wikipedia and Encyclopedia Britannica by Reagle and Rhue [RR11] by performing a manual analysis of biographies. In contrast, we follow a

computational linguistics approach where we do large-scale analysis of biographies of men and women in Wikipedia.

2.2.2 *Political Centralization*

In terms of systemic biases we study *political centralization*. Arguably, centralization can be considered an organizational schema instead of a systemic bias, yet in some developing countries this organization tends to favor the most central locations by making public policy favor its needs [Bra99] and concentrating economical power [GK08]. This systemic bias is widely discussed by Kollman [Kol13], who argues that, while centralization is not inherently bad, *"over-centralization is often irreversible and hard to avoid"*.

According to Gillespie and Robins [GR89], the technologies of communication that are supposed to shrink distances between communities are having the opposite effect, by constituting *"new and enhanced forms of inequality and uneven development"*. This is relevant in our context, as we work with Web platforms that are supposedly empowering users by removing physical barriers.

Dissertation Context

In Chapter 4 we study whether political centralization is reflected on Twitter and how this affects Web technology. We analyze user-generated content by the Chilean virtual population, which is centralized in the physical world [Lug08].

2.2.3 *Selective Exposure / Confirmation Bias and Homophily*

Festinger [Fes62] proposed the *cognitive dissonance* theory. Cognitive dissonance is a state of mental discomfort that appears in the presence of information that challenges current beliefs. In our context, when information seekers are presented with challenging and agreeable information side by side, they tend to choose the agreeable information even if it is not correct or it is less accurate than the challenging one. This is called *selective exposure* to information [Har+09]. A related bias, and sometimes used as synonym for selective exposure, is *confirmation bias* [Nic98], where people tend to seek information that

reinforces their beliefs. In both cases, an implicit purpose of such biased behavior is to avoid cognitive dissonance.

Another important bias in our context is *homophily*, originally defined by Lazarsfeld, Merton, *et al.* [L+54] and extensively reviewed by McPherson, Smith-Lovin, and Cook [MSC01]. It is the tendency to form ties with similar others, where similarity is related to beliefs, race, gender, and socio-demographic factors, among others. Homophily *per se* is not bad, as it makes easier to communicate with others, given common cultural backgrounds [Mar03].

Lazarsfeld, Merton, *et al.* [L+54] defined two types of homophily: *status* and *value*. Status homophily refers to those bonds created by social status characteristics, like education, ethnicity and age. Value homophily is the one of interest for this dissertation, as it refers to beliefs and think-alike ties. In political settings, establishing ties only with think-alike people causes *group polarization* [ML75], which, in turn, makes people in the group to make their points of view more extreme [Sun09].

Dissertation Context

In Chapter 5 we study whether *value homophily* is present in Twitter, and whether we can exploit partial homophily by finding people who are similar in non-political (or non-challenging) aspects.

2.3 WEB MINING, INFORMATION RETRIEVAL AND MACHINE LEARNING

Web Mining [Sri+00] is the discipline that focuses on finding and discovering patterns of behavior on the Web using *Data Mining* techniques over logged data and clickstreams. Techniques from fields like *Information Retrieval* [BR11a] and *Machine Learning* [Bis+06] complement the Web Mining process, described in Figure 2.3. This process contains three main stages: *pre-processing*, *pattern discovery* and *pattern analysis*, as well as an outcome of results that informs the implementation and design of adaptive Web applications [BKN07].

In the pre-processing stage, logged interaction data is cleaned and segmented into *user sessions* [CMS99]. Additionally, links between documents are used to build networks, and document content is represented in vectors that allow to perform mathematical operations with them (*e. g.*, computing similarity of two

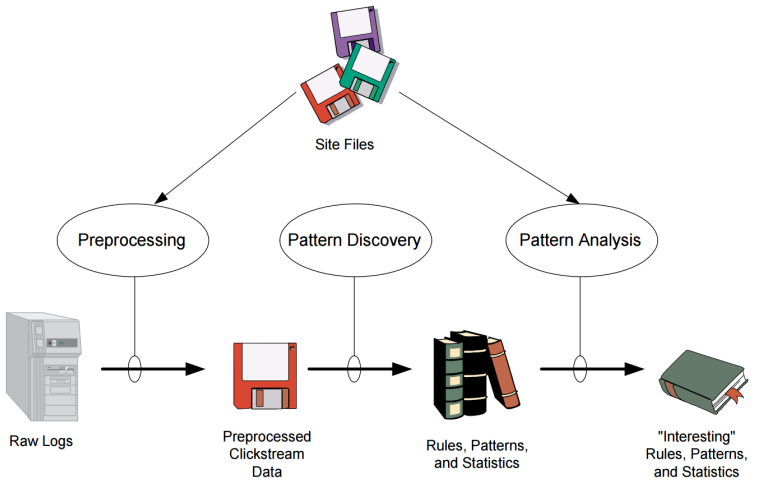


Figure 2.3: The Web Mining process by Srivastava *et al.* [Sri+00].

documents). Then, in the pattern discovery stage, several tasks are performed to find latent and explicit patterns in the data: *anomaly detection*, *association rules learning*, *clustering*, *classification*, *regression* and *summarization* [FPS96]. Finally, in the pattern analysis stage, the obtained results are validated in terms of their interestingness (either by a exploratory qualitative approach, or by quantitative approaches) and statistical hypothesis testing.

Note that this process is not necessarily linear. Exploratory, visual approaches exists, such as the *WET* visual analytics system by Pascual Cid [Pas10], which allows to interactively perform Web Mining tasks while visualizing intermediate results of analysis.

2.3.1 *Interaction Data: Server Logs and Clickstreams*

This is the core source of information for Web Mining. Clickstreams consist of events performed by users when they click on links on a page or by issuing search queries. An event contains the following information:

- *IP address* of the requester.
- *User-Agent* used, including name, version number and operating system.

- *Requested URL*.
- *HTTP Referer*: the URL that linked or preceded the current request.
- *Timestamp*: the time and date of the request.
- *Form Data*: any data attached to the request.

To identify sessions, IP address, User-Agent and Timestamp can be used. Navigation graphs can be made based on requested URLs and HTTP Referers. These graphs can be used to characterize navigation [Ben+09], as well as performing *query mining* [Bae05] in cases where the referer URL is from a search engine.

Dissertation Context

We apply Web Mining techniques to identify and quantify biases on user generated content. When analyzing interaction data from specific platforms like Twitter and Wikipedia some event data may be missing. For instance, microposts do not contain IP address nor Request URLs, but they include meta-data (*e. g.*, user entities, links, hashtags, and possibly geographical coordinates) useful to identify and characterize users. In addition, we analyze logged data from end-user interaction with the implemented systems in Chapters 4 and 5.

2.3.2 *Networks on the Web*

Hyperlinks on the Web allow the creation of networks based on them. Barabási and Albert [BA99] found that these networks are *scale-free*, which means that degree distribution follows a *powerlaw*. These networks are characterized by the presence of *hubs*, nodes with much higher degree than the other nodes, as shown on Figure 2.4. These distributions have been found in subsets of the Web [BCE07], including the platforms under study in this dissertation: Twitter's follower graph [Kwa+10] and Wikipedia's network of links between articles [Zla+06].

Networks are used when analyzing patterns and providing input for algorithms to generate adaptive content, as well as to estimate node importance. Two common algorithms to perform such estimation are *PageRank* [Pag+99] and *HITS* [Kle99]. Twitter's recommender system uses an algorithm inspired by both named *SALSA* [Gup+13], and PageRank has been used to estimate historical importance of Wikipedia entities [Ara+12].

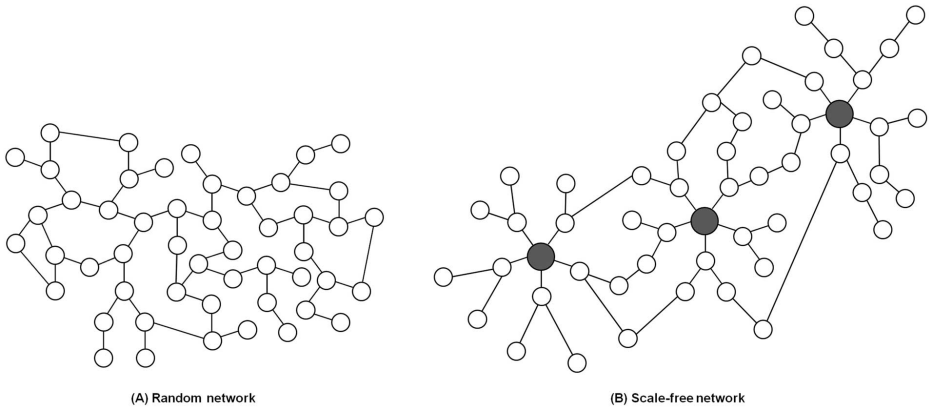


Figure 2.4: Random (Left) and Scale-Free Networks (Right). Image by Seo *et al.* [Seo+13].

Dissertation Context

In Chapter 3 we use PageRank over a graph of links between biography articles of Wikipedia, in a similar way to Aragón *et al.* [Ara+12]. In Chapters 4 and 5 we build networks based on how Twitter users interact with others, and analyze those networks using centrality metrics described next.

2.4 SOCIAL SCIENCES AND SOCIAL NETWORK ANALYSIS

We refer to Social Sciences as those disciplines concerned with human behavior. There are many disciplines that fall under Social Sciences, like Sociology [Gid+00], Behavioral Science, Linguistics, Psychology and Anthropology. In particular, Sociology is the study of social behavior, from micro (individuals) to macro (institutions and systems) levels.

Social behavior can be represented in graphs, which are analyzed with Social Network Analysis (SNA) tools. SNA is the use of network theory to study social networks, in terms of the connections in the network, the distribution of those connections, and their possible segmentation in communities [Fre04]. It is an interdisciplinary area, where sociological ideas and the computing power of Web Mining techniques allows to perform complex analysis over massive social networks.

2.4.1 *Centrality*

Centrality is a measure of importance for nodes in graphs, based on the link structure of the graph and the degree of each node. Some measures include: *betweenness centrality* [Fre77], defined as the fraction of shortest paths that include each node; *closeness centrality* [BF05], defined as the sum of distances from a node to all other nodes; and *eigenvector centrality*, which assigns scores to all nodes in the network that depend on the scores of the connections of each one, meaning that connecting to higher-score nodes implies a higher-score. PageRank [Pag+99] is a variant of eigenvector centrality.

Dissertation Context

In Chapter 3 we quantify bias on network structure by comparing the observed network with several *null models*, with centrality estimated with PageRank [Pag+99]. In Chapter 4 we use *random-walk betweenness centrality* [New05] to evaluate if political centralization is reflected from the physical world on the virtual population in Twitter. In Chapter 5 we use *information centrality* [BF05] to find topics with potential to connect people of opposing views.

2.4.2 *Small Worlds*

Milgram [Mil67] asked whether we live in a small world or not, given the *six degrees of separation* premise. This concept has been adopted in Social Network Analysis, where small-world networks are defined as those where most nodes are not neighbors of one another, but the shortest distances between nodes are small. Watts and Strogatz [WS98] defined a specific case of small world where, in addition to small node distances, nodes exhibit a high clustering coefficient. Figure 2.5 displays differences between regular, random, and small world networks.

Mislove *et al.* [Mis+07] confirmed the scale-free, small world structure of many Web platforms, including Twitter.

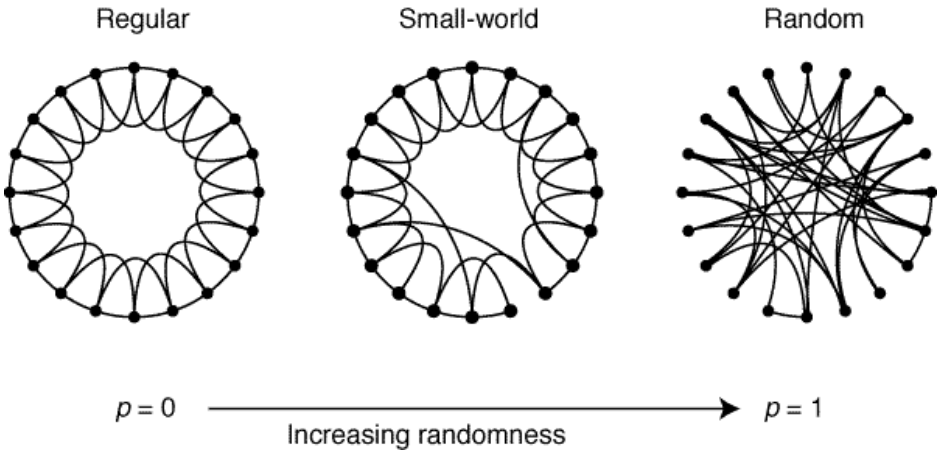


Figure 2.5: Small Worlds in comparison with Regular and Random Networks, by Watts and Strogatz [WS98].

Dissertation Context

In Chapter 3, we emphasize bias on network structure by comparing the observed network with several *null models*. One of them is the small-world model by Watts and Strogatz [WS98].

2.4.3 *The Self and the Other(s)*

From a qualitative point of view, we address self-presentation as defined by Goffman [Gof59] as well as identity building, in particular from "the other". We consider "the other" from three perspectives. First, according to Simone de Beauvoir [De 12], as a way to differentiating men and women, when women are the others, as they are *"the minority"*. Second, we focus on an ethnographic view on group identification, which, according to Butler [But06], is based on the differences with the others: *"The one with whom I identify is not me, and that 'not being me' is the condition of the identification"*. Third, when avoiding connecting with others who do not think-alike because of their challenging context.

Dissertation Context

In Chapter 3, we analyze the results of the bias quantification in terms of social theories exposed by feminist authors. In Chapter 4 we learn that identification as defined by Butler [But06] is a key concept to consider when designing user interfaces to encourage exploration of geographically diverse informational content. Finally, in Chapter 5 we consider the idea of self-presentation by Goffman [Gof59] to reinforce user engagement with the proposed interface design.

2.5 INFORMATION VISUALIZATION

As Bederson and Shneiderman [BS03] indicate, Information Visualization is an interdisciplinary field emerged *"from research in human-computer interaction, computer science, graphics, visual design, psychology, and business methods"*. Heer, Bostock, and Ogievetsky [HBO10] indicate that its goal is to ease understanding of data by leveraging human cognition to see patterns, trends, and identify outliers from data. They provide a survey of common techniques and their corresponding data types. These data types range from *multidimensional*, *temporal*, *network*, or *tree* data [Shn96]. As example, Figure 2.6 displays one of the most famous examples of visualization, done by Charles Minard in 1869 to depict the march of Napoleon's troops to Russia in the winter of 1812. It is a flow diagram over a map that illustrates the size of the army during advance and then retreat from Moscow. The diagram includes a time-series that allows readers to see temperature and link its value with the progress of the trip made by the troops, including the geographical context.

Particularly, the visualization by Charles Minard is referred as an *information graphic*. Not all visualizations are equal, nor with the same purposes. The following are some subfields or research areas in Information Visualization:

- *Visual Analytics and Visual Data Mining* [Kei+08], where visualization allows users to enhance data mining processes which, automated, would not be as powerful as the synergy between human intervention (through visualization) and data mining.
- *Casual Information Visualization* [PSM07], where visualizations are not task-based, and users are expected to be non-experts. Yet, even without

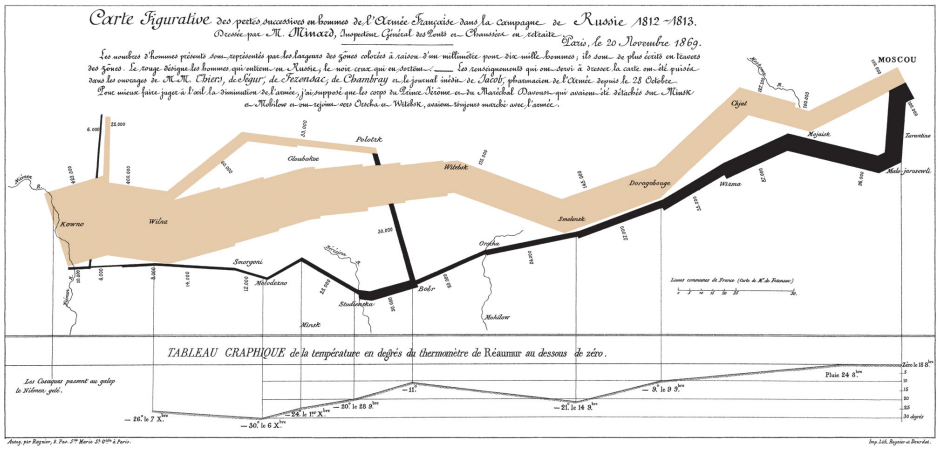


Figure 2.6: Charles Minard's map of Napoleon's Russian campaign of 1812, made in 1869. Source: Friendly [Fri02].

tasks nor expertise, different kinds of insights can be obtained, like *social insights*.

- *Social Data Analysis* [WK06], where visualization is used as a tool to reach a wider audience who will analyze their own social data.
- *Information Graphics* [Cai12], where visualization techniques are used to create graphical depictions of data to tell stories, usually in journalist settings.

For a survey of common visualization techniques see Heer, Bostock, and Ogievetsky [HBO10]. Research-wise, Liu *et al.* [Liu+14] provides a comprehensive survey of recent techniques and advances.

2.5.1 *Visualization Design: From Data to Information*

When designing visualizations, regardless of the specific area (if any) being targeted, the *mental model* of the target users of the system must be considered. Moreover, Liu and Stasko [LS10] proposes that visualization internalization follows a process of four stages: *internalization*, *processing*, *augmentation*, and *creation*. Thus, insights are not expected to appear immediately when interacting with/being exposed to visualizations. This cognitive view of visualization

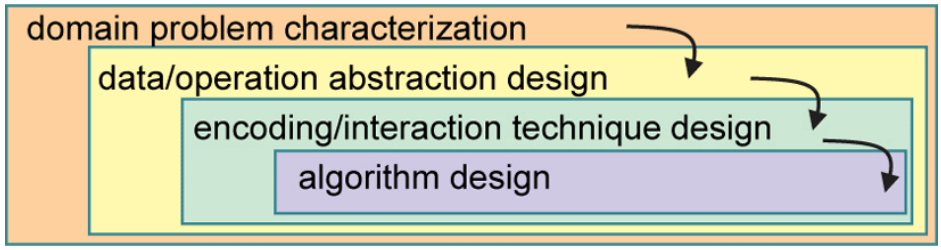


Figure 2.7: Nested visualization design and validation model by Munzner [Mun09].

is compatible with a technologist view, where visualization is thought as a *technology tool* [Cai12]. Although there are known visualization techniques that are more suitable for some data types than others [HBO10], there must be a *design process* behind. Munzner [Mun09] proposes a nested model with four stages: *characterize the task and data in the vocabulary of the problem domain*, *abstract into operations and data types*, *design visual encoding and interaction techniques*, and *create algorithms to execute techniques efficiently*.

In the nested model by Munzner [Mun09], *domain problem characterization* is at the root. Yet, as Van Wijk [Van06] asks, how does the designer approach the gap between her/him and the target users? In the Information Visualization process defined by Dürsteler and Engelhardt [DE07], the visualization designer must consider the cultural and social context between her/him and the user, as displayed on Figure 2.8. These contexts must be considered when designing *user experiences* [Bux10], as they mold the way a user thinks and acts with information.

To characterize visual designs, Cairo [Cai12] defined the *visualization wheel*, displayed on Figure 2.9. In this framework, several pairs of attributes are defined: *abstraction* and *figuration*, *functionality* and *decoration*, *density* and *lightness*, *multi-* and *uni-dimensionality*, *originality* and *familiarity*, and *novelty* and *redundancy*. Each pair must be balanced when planning and designing a visualization. On each pair, a tendency to the first one produces a deeper and complex design; conversely, a tendency to the second one produces a more intelligible and shallower design.

The final step we consider, which is not included in the process defined by Dürsteler and Engelhardt [DE07], but is considered in the model by Munzner

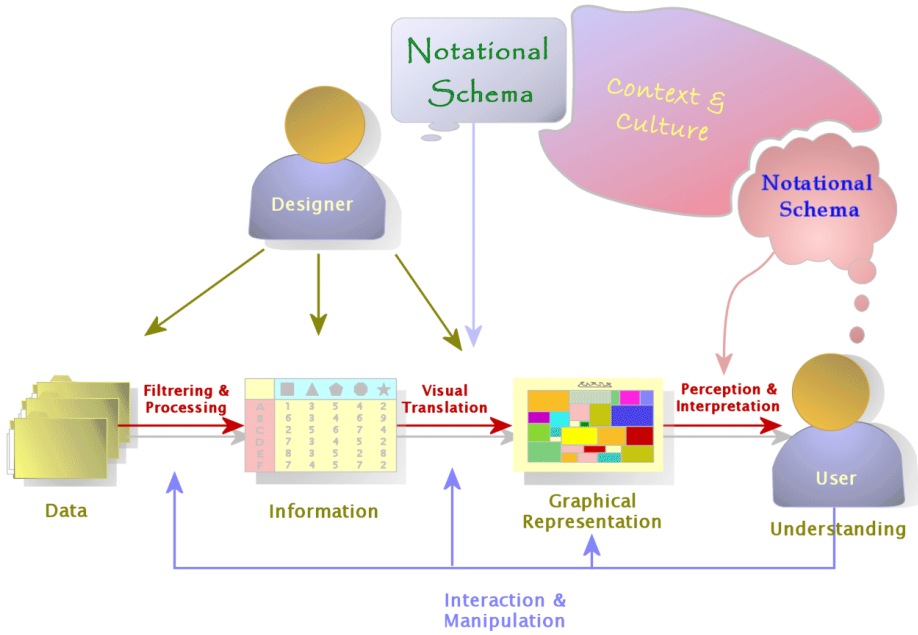


Figure 2.8: The process of Information Visualization by Dürsteler and Engelhardt [DE07].

[Mun09], is the evaluation and validation of the design. Lam *et al.* [Lam+12] surveyed more than three hundred papers from visualization-related venues, and devised seven scenarios for evaluation, based on *process* and *visualization*:

- *Process*: understanding environments and work practices, evaluating visual data analysis and reasoning, evaluating communication through visualization, and evaluating collaborative data analysis.
- *Visualization*: evaluating user performance, evaluating user experience, and evaluating visualization algorithms.

Some evaluation techniques to consider in each scenario include: controlled experiments, qualitative methods (fields observations, interviews), case studies with domain experts, and interaction data analysis. For an in-depth analysis of each scenario we refer the reader to the original survey [Lam+12].

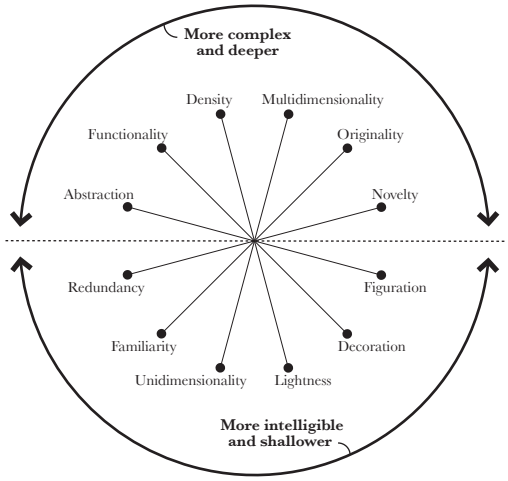


Figure 2.9: The *visualization wheel* by Cairo [Cai12].

Dissertation Context

In Chapters 4 and 5 we design user interfaces with visualization components where the design decisions were made considering these processes. We focus on more intelligible and shallower designs, as users of micro-blogging platforms are mostly not experts. Particularly, in Chapter 5 our work is clearly situated in the field of *Casual Information Visualization* [PSM07].

To evaluate our designs, we focus on the analysis of interaction data with Web-based implementations of them. This is known as "in the wild" evaluation [Cra+13], because the environment is not controlled, and neither are the kinds of users who access the system. Because our visualizations do not consider specific tasks, instead of evaluating user performance [Lam+12], we evaluate *user engagement* metrics [LOY14], in particular dwell time and the tendency of users to return to a site.

3

GENDER BIAS IN WIKIPEDIA

Contributing to history has never been as easy as it is today. Anyone with access to the Web is able to play a part on Wikipedia, an open and free encyclopedia. Wikipedia, available in many languages, is one of the most visited websites in the world and arguably one of the primary sources of knowledge on the Web. However, not *everyone* is contributing to Wikipedia from a diversity point of view; several groups are severely underrepresented. One of those groups is *women*, who make up approximately 16% of the current contributor community, meaning that most of the content is written by men. In addition, although there are specific guidelines of verifiability, notability, and neutral point of view that must be adhered by Wikipedia content, these guidelines are supervised and enforced by men.

In this chapter, we propose that gender bias is not about participation and representation only, but also about characterization of women. We approach the analysis of gender bias by defining a methodology for comparing the characterizations of men and women in biographies in three aspects: meta-data, language, and network structure. Our results show that, indeed, there are differences in characterization and structure. Some of these differences are reflected from the off-line world documented by Wikipedia, but other differences can be attributed to gender bias in Wikipedia content. We contextualize these differences in feminist theory and discuss their implications for Wikipedia policy.

3.1 INTRODUCTION

Today's Web creates opportunities for global and democratic media, where everyone has a voice. One of the most visible examples is Wikipedia, an open encyclopedia where anyone can contribute content. In contrast to traditional encyclopedias, where a staff of experts in specific areas takes care of writing, editing and validating content, in Wikipedia these tasks are performed by a community of volunteers. Whether or not this *open source* approach provides reliable and accurate content [Gil05; Ros06], Wikipedia has gained unprecedented reach. Indeed, Wikipedia was the 7th most visited website during 2014.[1] An extensive body of research builds upon Wikipedia [Oko+14], covering topics like participation, structured data, and analysis of historical figures, among others.

In theory, by following its guidelines about verifiability, notability, and neutral point of view, Wikipedia should be an unbiased source of knowledge. In practice, the community of Wikipedians is not diverse, and contributors are inherently biased. One group that is severely underrepresented in Wikipedia is women, who represent only 16% of editors [HS13]. This disparity has been called the *gender gap* in Wikipedia, and has been studied from several perspectives to understand why more women do not join Wikipedia, and what can be done about it. It is a problem because reportedly women are not being treated as equals to men in the community [Lam+11], and potentially, in content. For instance, Filipacchi [Fil13] described a controversy where women novelists started to be excluded from the category *"American Novelists"* to be included in the specific category *"American Women Novelists."*

Instead of focusing on the participatory *gender gap*, we focus on how women are characterized in Wikipedia articles, to assess whether gender bias from the off-line world extends to Wikipedia content, and to identify biases exhibited by Wikipedians in the characterization of women and of their historical significance. The research questions that drive our work are:

> *Is there a gender bias in user-generated characterizations of men and women in Wikipedia?*
> *If so, how to identify and quantify it? How to explain it based on social theory?*

1 http://www.alexa.com/siteinfo/wikipedia.org

The study of biases in Wikipedia is not new. Hecht and Gergle [HG09] defined the notion of *self-focus bias* to study the cultural biases present in Wikipedia from a "hyperlingual" approach. Having the gender gap in mind, we focus on gender bias not only to quantify it, but to understand what could be causing it. As a first approach to the problem, we focus on the English language to be able to analyze our results in terms of western feminist theories from the social sciences.

In the book *The Second Sex*, Simone de Beauvoir widely discusses different aspects of women oppression and their historical significance. She wrote in 1949: *"it is not women's inferiority that has determined their historical insignificance: it is their historical insignificance that has doomed them to inferiority"* [De 12]. More than 60 years later, almost anyone with access to the Web can contribute to the writing of history, thanks to Wikipedia. The scale of Wikipedia, as well as its openness, allows us to perform a quantitative analysis of how women are characterized in Wikipedia in comparison to men. Encyclopedias characterize men and women in many ways, *e. g.*, in terms of their lives and the events in which they participated or were relevant. We concentrate on *biographies* because they are a good source to study gender bias, given that each article is about a specific person.We propose three dimensions along which to perform our analysis: *meta-data*, *language*, and *network structure*. This leads to three major findings:

1. Differences in meta-data are coherent with results in previous work, where women biographies were found to contain more marriage-related events than men's.
2. Sex-related content is more frequent in women biographies than men's, while cognition-related content is more highlighted in men biographies than women's.
3. A strong bias in the linking patterns results in a network structure in which articles about men are disproportionately more central than articles about women.

The main contributions of this work are methods to quantify gender bias in user generated content, a contextualization of differences found in terms of feminist theory, and a discussion of the implications of our findings for informing policy design in Wikipedia. As said earlier, we focus on the English Wikipedia, but our methods are generalizable to other languages and platforms.

3.2 BACKGROUND

Research on the community structure and evolution of Wikipedia has been prominent. In its first steps, the focus was on growth [AMC07] and dynamics [Rat+10], without attention toward gender. Later, it was found that there is a gender gap, as Wikipedia has fewer contributions from women, and women stop contributing earlier than men [Lam+11]. There are differences in how genders behave. For instance, men and women communicate differently in the inner communication channels in Wikipedia [Lan+12]: they focus on different topics [Lam+11] and the level of content revision differs by gender but also by amount of activity [Ant+11]. In addition, Lam *et al.* [Lam+11] found that women are more *reverted* than men (*i.e.*, their contributions are discarded), and reportedly women contribute less because of aggressive behavior toward them [CB12; Sti13]. Efforts have been made to build a more welcoming community and to encourage participation [Mor+13; CT14], and Wikimedia itself encourages initiatives like *WikiWomen's Collaborative*.[2]

Content-wise, the study of biographies in Wikipedia enables cultural comparisons of coverage [CH11], as well as the construction of social networks of historical (and current) figures [Ara+12]. Although bias in content has been addressed before through *self-focus bias* [HG09], such bias has been measured at large-scale only in terms of culture, not gender. Lam *et al.* [Lam+11] found that coverage of "female topics" was inferior to "male topics " when classifying topics as "male" or "female" according to the people who contributed to them. Reagle and Rhue [RR11] found that in characterization of women, in comparison to commercial encyclopedias like *Britannica*, Wikipedia has better coverage of notable profiles, although this coverage is quite low and it is still biased towards men. Bamman and Smith [BS14] found that women biographies are more likely to include marriage or divorce events.

Addressing the gender gap from a content perspective may help to improve the quality and value of the content. Currently, focus on quality in Wikipedia has been about predicting article quality [ASL12; FFG14]. However, focusing on quality without considering readers does not give the whole picture, as Wikipedia readers are not necessarily interested in the same topics as contribu-

2 http://meta.wikimedia.org/wiki/WikiWomen's_Collaborative

tors [Leh+14a] and might have a different concept of quality. Moreover, in our context, Flekova, Ferschke, and Gurevych [FFG14] found that quality of biographies is assessed differently depending on the gender of the portrayed person. Is it because the raters were biased? Or is it because biographies were written differently? Our hypothesis is that biographies are written differently, an idea inspired by seminal work about how women are characterized by language [Lak73].

To study differences in text, word frequency is commonly used. Word frequency follows Zipf's law [Zip49; SFM09], an empirical distribution found in many languages [Pia14]. An interesting property of Zipf distributions in language is that small sets of words that are semantically or categorically related also follow a Zipf distribution [Pia14]. This property implies that, given two subsets of words that are related semantically or categorically, their frequency distributions can be compared. Thus, we compare frequency distributions according to gender for several semantic categories derived from the *Linguistic Inquiry and Word Count* (LIWC) dictionary. LIWC studies *"emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples"* [PFB01]. It has been used to analyze interactions between Wikipedia contributors [Ios+14] and article content with respect to emotions [FM12]. In a context similar to ours, Schmader, Whitehead, and Wysocki [SWW07] used LIWC to quantify differences in characterization of women and men in recommendation letters.

In our work, we quantify gender bias in Wikipedia's characterization of men and women through their biographies. To do so we approach three different dimensions of biographies, which we analyze in different sections on this chapter: *meta-data*, provided by the structured version of Wikipedia, DBPedia [Leh+14b]; *language*, considering how frequent are words and concepts [SFM09]; and *network structure*. In terms of network structure, we build a biography network [Ara+12] in which we estimate PageRank, a measure of node centrality based on network connectivity [BP98; For+07]. In similar contexts, PageRank has been used to provide an approximation of historical importance [Ara+12; SW14] and to study the bias leading to the gender gap [SW14]. We measure bias in link formation by comparing the importance given by PageRank in the biography network with those of null models, *i. e.*, graphs that are unbiased by construction but that maintain certain properties of the source biography network.

Figure 3.1: *Infobox* from the biography article of Simone de Beauvoir.

3.3 DATASET AND META-DATA PROPERTIES

To study gender bias in Wikipedia, we consider three freely available data sources:

1. The DBPedia 2014 dataset [Leh+14b].[3]
2. The Wikipedia English Dump of October 2014.[4]
3. Inferred gender for Wikipedia biographies by Bamman and Smith [BS14].[5]

3 http://wiki.dbpedia.org/Downloads2014
4 https://dumps.wikimedia.org/enwiki/20141008/
5 http://www.ark.cs.cmu.edu/bio/

DBPedia and Meta-data

DBPedia is a structured version of Wikipedia that provides meta-data for articles, normalized article URIs (*Uniform Resource Identifiers*), and normalized links between articles (taking care of redirections). It provides a shallow hierarchy of classes, which includes a *Person* category. To provide the structured meta-data, DBPedia processes from the content of infoboxes in Wikipedia articles. Infoboxes are template-based specifications for specific kinds of articles. When DBPedia detects an infobox with a template that matches those of a person, it assigns the article to the *Person* class, and to a specific subclass if applicable (*e. g.*, *Artist*). For instance, Figure 3.1 displays the infobox of *Simone de Beauvoir*[6] [De 12]. The infobox contains specific meta-data pertinent to a biography, such as date and place of birth, but it does not include gender (in specific cases it does, see "Inferred Gender" next). DBPedia maps infobox properties to specific fields in a person's meta-data. These properties are not always available in the infobox templates, and do not always have a standardized name. DBPedia, whenever possible, normalizes both attribute keys and attribute values.

Wikipedia Biographies

We consider two versions of the biographies: the overview and the full text. We analyze both in different contexts: in the overview we analyze the full vocabulary employed, while in the full text we analyze only the words pertaining to the LIWC dictionaries. The overview is described by Wikipedia as *"an introduction to the article and a summary of its most important aspects. It should be able to stand alone as a concise overview."* Since those aspects are subjective, the introduction content is a good proxy for any potential biases expressed by Wikipedia contributors. At the same time we avoid potential noise included in the full biography text from elements like quotations and the filmography of a given actor/actress. In both cases (overview and full content), template markup is removed from analysis.

6 https://en.wikipedia.org/wiki/Simone_de_Beauvoir

Inferred Gender

To obtain gender meta-data for biographies, we match article URIs with the dataset by Bamman and Smith [BS14], which contains inferred gender for biographies based on the number of grammatically gendered words (*i. e.*, *he*, *she*, *him*, *her*, etc.) present in the article text. Bamman and Smith [BS14] tested their method in a random set of 500 biographies, providing 100% precision and 97.6% recall. This method has also been used before by Reagle and Rhue [RR11] and DBPedia itself [Leh+14b], making DBPedia to include gender meta-data in some cases. However, note that the genders considered in these datasets (and thus, in this work) are only *male* and *female*.

3.3.1 *Meta-Data Properties*

In our first analysis we estimate the proportion of women in Wikipedia. We analyze meta-data by comparing how men and women proportionally have several attributes in the data from DBPedia.

Presence and Proportion According to Class

DBPedia estimates the length (in characters) and provides the connectivity of articles. Of the set of 1,445,021 biographies (articles in the DBPedia Person class), 893,380 (61.82%) have gender meta-data. Of those, only 15.5% are about women.

The mean article length is 5,955 characters for men and 6,013 characters for women (a significant difference according to a t-test for independent samples: $p < 0.01$, Cohen's $d = 0.01$). The mean out-degrees (number of links) of 42.1 for men and 39.4 for women also differ significantly ($p < 0.001$, Cohen's $d = 0.06$). Table 3.1 displays the number of biographies in the *Person* class, as well as its most common subclasses with their corresponding out-degrees per gender. From the table, in comparison to the global proportion of women, the following categories over-represent women: *Artist*, *Royalty*, *FictionalCharacter*, *Noble*, *BeautyQueen*, and *Model*. The others over-represent men. The differences in length and degree do not hold for all classes, hinting that a study according to semantic categories of people is needed. However, in this chapter we focus on the global differences in *Person*.

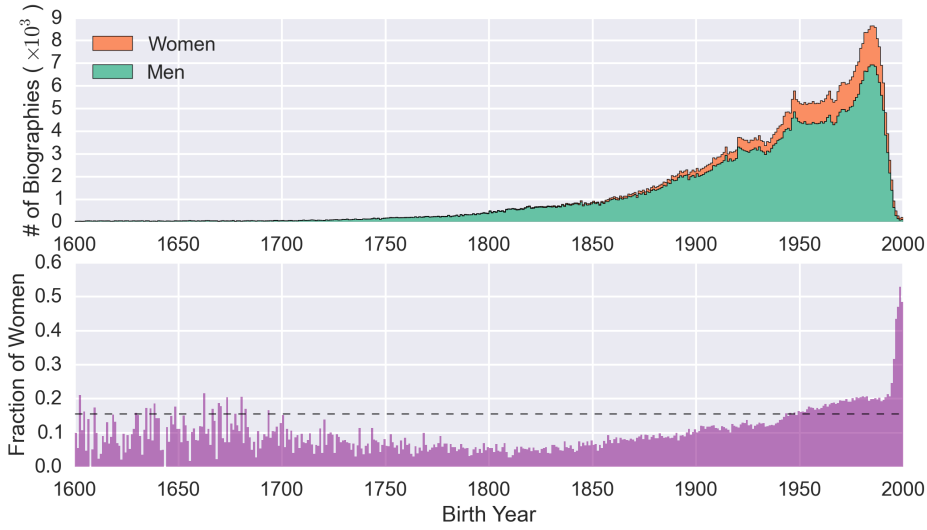


Figure 3.2: Distribution of biographies according to birth year.

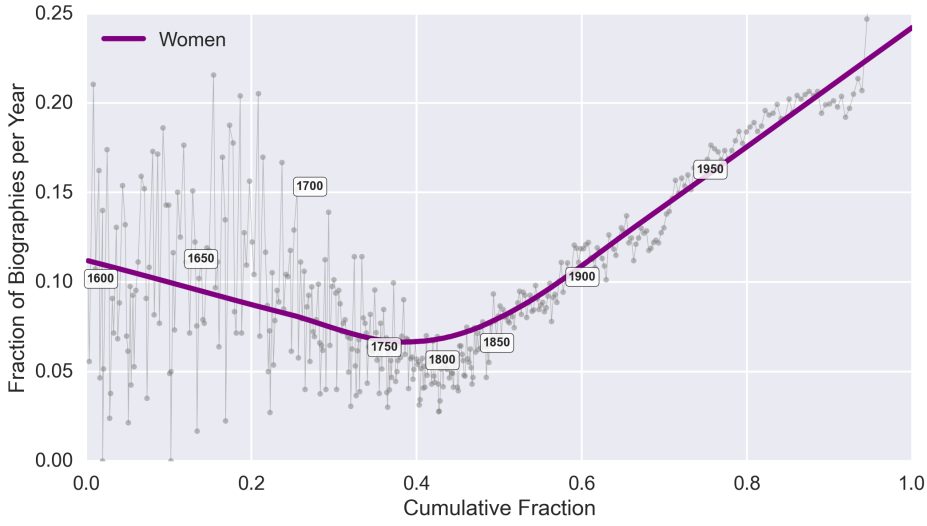


Figure 3.3: Relation between the cumulative fraction of women and the fraction of women per year (dots). The y-axis was truncated to 0.25 for clarity.

Distribution According to Date of Birth

Figure 3.2 displays the distribution of biographies according to their corresponding *birthYear* property, considering only those biographies between years 1600 and 2000 (inclusive). This accounts for 65.48% of biographies with gender (note that 34.07% does not have date of birth in meta-data). The distribution per gender (top chart) shows that most of the biographies of both genders are about people from modern times. The distribution of fraction of women per year (bottom chart) shows that since the year 1943 the fraction of women is consistently above the global value of 0.155. Note that, of the biographies that have date of birth in their meta-data, 53% are from 1943 until 2000. To explore the evolution of growth of women presence, in Figure 3.3 we display the relationship between the cumulative fraction of biographies and the yearly fraction of biographies of women. The chart includes a *LOWESS*[7] fit of the data, to be able to see the tendency of changes in representation. This tendency became positive in the period 1750–1800. These results are discussed in terms of historical significance in the discussion section.

Infobox Attributes

Given that there are different classes of infoboxes, there are many different meta-data attributes than can be included in biographies. In total, we identified 340 attributes. For each one of them, we counted the number of biographies that contained it, and then compared the relative proportions between genders with a chi-square test. Only 3.53% presented statistically significant differences. Those attributes are displayed in Table 3.2. All of them have large effect sizes (Cohen's $w > 0.5$). Inspection allows us to make several observations:

- Attributes *careerStation*, *formerTeam*, *numberOfMatches*, *position*, *team*, and *years* are more frequent in men. All these attributes are related to sports, and thus, these differences can be explained by of the prominence of men in sports-related classes (*e. g.*, *Athlete*, *SportsManager* and *Coach* in Table 3.1).

7 Locally weighted scatterplot smoothing.

Table 3.1: Number of biographies in the dataset for the Person class and its most common child classes (in terms of biographies with gender). In this and the following tables, we use this legend for p-values: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Ontology	With gender	% Women	M. OutD.	W. OutD.	OutD. t	M. Len.	W. Len.	Len. t
Person	893380	15.53	42.07	39.39	20.77***	5955	6013	-2.65**
Athlete	187828	8.94	50.28	45.64	10.64***	6203	6383	-2.83**
Artist	79690	25.14	53.02	47.06	12.95***	7670	7695	-0.33
OfficeHolder	38111	13.04	53.67	45.23	10.97***	8369	7732	3.77***
Politician	32398	8.75	42.74	41.73	1.29	6026	6668	-4.02***
MilitaryPerson	22769	1.67	61.41	52.41	4***	8269	7818	1.03
Scientist	15853	8.79	49.12	43.66	4.91***	8111	8115	-0.01
SportsManager	11255	0.62	64.92	59.94	0.79	7090	9663	-2.79**
Cleric	8949	6.34	46.68	41.06	3.23**	6324	6316	0.02
Royalty	7054	35.24	68.83	67.98	0.55	9294	8800	1.75
Coach	5720	2.40	49.08	48.09	0.27	8318	10055	-2.65**
FictionalCharacter	4023	26.08	62.72	56.24	3.03**	12256	12063	0.39
Noble	3696	23.16	46.34	42.44	3.16**	5863	5439	2.05*
Criminal	1976	12.45	51.35	48.41	1.08	11781	13282	-1.69
Judge	1949	14.88	43.24	34.53	3.93***	6502	5014	2.97**
Monarch	1861	6.23	61	41.45	3.40***	9673	6258	2.91**
Architect	1730	3.76	52.96	30.77	3.29**	8135	5400	2.17*
BeautyQueen	1464	99.59	26	33.06	-0.70	3374	4074	-0.45
Philosopher	1304	7.13	78.95	62.68	2.04*	15319	12177	1.66
Model	1267	89.34	34.57	40.20	-2.01*	5139	6205	-1.90

- Attributes *deathDate*, *deathYear* are more frequent in men. According to Figure 3.2, most women are from recent times, and thus they are presumably still alive.
- Attribute *birthName* is more frequent in women. Its values refer mostly to the original name of artists, and women have considerable presence in this class (see Table 3.1). In addition, other possible explanation is that, in the case of married women, they usually change their surnames to those of their husbands.
- Attributes *occupation* and *title* are more frequent in women. *Title* is a description of a person's occupation (the most common are *Actor* and *Actress*), while *occupation* is a DBPedia resource URI (*e.g.*, http://dbpedia.org/resource/Actor). The infoboxes of sport-related biographies do not contain these attributes because their templates are already indicators of their occupations, and thus, athletes (which are mostly men) do not contain such attributes.

The case of the *spouse* attribute is different. The inspection does not offer a direct explanation other than the tendency to include this attribute more in women biographies than in men's. For instance, the most common class with the spouse attribute is *Person*, the reference class, with 45% of the instances of the attribute.

3.4 LEXICAL PROPERTIES

In this section we explore the characterization of women and men from a lexical perspective. We analyze the vocabulary used in the overview of each biography through word frequency, and we use the estimated frequencies to find which words are associated with each gender. To estimate relative frequencies, words were considered once per biography, and we estimated bi-gram word collocations to identify composite concepts (*e. g.*, *New York*). We obtained a vocabulary of size $V_m = 1{,}013{,}305$ for men, $V_w = 376{,}737$ for women, with $V = 272{,}006$ common words.

Figure 3.4 displays a density plot of word frequency, and the Probability Density Functions (PDFs) for both genders. The frequency distributions are similar across genders. Word frequencies in the common vocabulary for both gen-

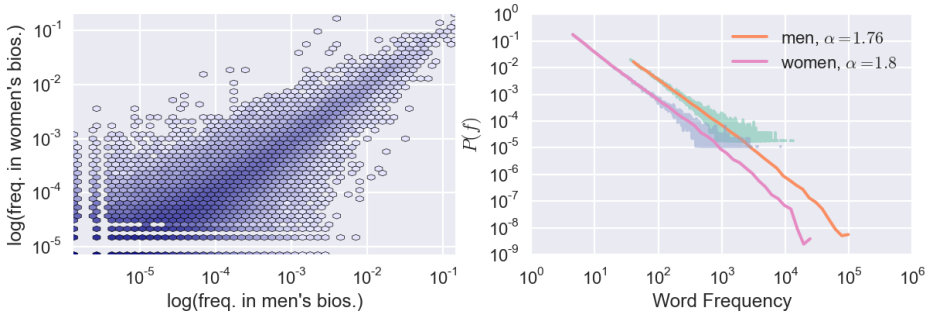


Figure 3.4: A density hexbin plot of word frequencies in men/women's biographies (left), and the PDF of word frequency distribution according to gender (right). Fitting to Zipfian distributions with the *powerlaw* library [ABP14] yields the shown exponents.

Figure 3.5: Words most associated with women (left) and men (right), estimated with *Pointwise Mutual Information*. Font size is inversely proportional to PMI rank. Color encodes frequency (the darker, the more frequent).

Table 3.2: Proportion of men and women who have the specified attributes in their infoboxes. Proportions were tested with a chi-square test, with effect size estimated using Cohen's w.

	% Men	% Women	χ^2	w
birthName	4.01	11.46	4.84^{*}	0.81
careerStation	8.95	1.13	6.84^{**}	0.94
deathDate	32.82	19.35	5.53^{*}	0.64
deathYear	44.68	25.45	8.28^{**}	0.66
formerTeam	4.40	0.24	3.94^{*}	0.97
numberOfMatches	8.60	1.06	6.61^{*}	0.94
occupation	12.52	23.28	4.97^{*}	0.68
position	13.62	1.68	10.46^{**}	0.94
spouse	1.56	6.86	4.10^{*}	0.88
team	14.06	1.97	10.39^{**}	0.93
title	9.17	19.65	5.59^{*}	0.73
years	8.95	1.12	6.84^{**}	0.94

ders follow a Zipf distribution $P(f) \sim f^{-\alpha}$ with similar exponents $\alpha \approx 1.8$, consistent with the value found by Serrano, Flammini, and Menczer [SFM09]. In addition, frequency with respect to gender presents a high rank-correlation $\rho = 0.65$ ($p < 0.001$). For reference, consider that the inter-language rank correlation of words with the same meaning across languages is 0.54 [CP11]. This implies that words share meanings when referring to men and women.

3.4.1 *Associativity of Words with Gender*

To explore which words are more strongly associated with each gender, we measure *Pointwise Mutual Information* [CH90] over the set of vocabulary in both genders. PMI is defined as:

$$\mathrm{PMI}(c, w) = \log \frac{p(c, w)}{p(c)p(w)}$$

where c is a class (*men* or *women*), and w is a word. The probabilities can be estimated from the proportions of biographies about men and women, and the corresponding proportions of words and bi-grams. Since PMI overweights words

with very small frequencies, we consider only words that appear in at least 1% of men or women biographies.

Associativity results are displayed as word clouds in Figure 3.5. The top-15 words associated to each gender are (relative frequency in parentheses):

- Women: *actress* (15.9%), *women's* (8.8%), *female* (5.6%), *her husband* (4.1%), *women* (5.3%), *first woman* (1.9%), *film actress* (1.6%), *her mother* (1.8%), *woman* (4.4%), *nee* (3.6%), *feminist* (1%), *miss* (1.9%), *model* (3.3%), *girls* (1.5%) and *singer* (6.5%).
- Men: *played* (14.2%), *footballer who* (3.0%), *football* (4.5%), *league* (5.9%), *john* (7.9%), *major league* (1.8%), *football league* (1.6%), *college football* (1.5%), *son* (7%), *football player* (2.2%), *footballer* (2%), *served* (11.7%), *william* (4.6%), *national football* (2%) and *professional footballer* (1%).

Clearly, the words most associated with men are related to sports, football in particular, which refers to both popular sports of soccer and American football (recall from Table 3.1 that *Athlete* is the largest subclass of *Person*). For women, the most associated words are related to arts (recall from Table 3.1 that *Artist* is the second largest subclass of *Person*), gender (*women's*, *female*, *first woman*, *feminist*), and family roles (*her husband*, *her mother*, *nee*[8]). This is consistent with the results from the meta-data analysis, where women are more likely to have a *spouse* attribute in their infoboxes (see Table 3.2), and with the results of Bamman and Smith [BS14].

3.4.2 *Gender Differences in Semantic Categories of Words*

Words most associated to each gender might belong to categories that are hard to compare, given their richness and complexity. We use the *Linguistic Inquiry and Word Count* dictionary of semantic categories to find if different genders have different characterizations according to those semantic categories. The LIWC dictionary includes, for each category (and its corresponding subcategories), a list of words and prefixes that match relevant words. We consider the following pertinent categories to our context: *Social Processes*, *Cognitive Processes*, *Biological Processes*, *Work Concerns* and *Achievement Concerns*. To generate the final dictionaries from the vocabulary, we matched the prefixes in our

8 Adjective used when giving a former name of a woman.

corpus and performed manual cleaning of noisy keywords like place names (*e. g.*, *Virginia* matches *virgin** from the *sexual* category), surnames (*Lynch* matches *lynch** from the *death* category), and words with unrelated meanings. In total, our cleaned dictionary contained 2,877 words.

To compare the distribution of words in the semantic categories, we employed two metrics: relative frequency in overviews, as previously done with PMI, and burstiness in the full text. Word frequencies identify how language is used differently to characterize men and women in terms of semantic categories. However, word frequency alone does not give insights on how those semantic categories portray a given biography, or in other words, the importance that editors give to those categories. Burstiness is a measure of word importance in a single document according to the number of times it appears within the document, under the assumption that important words appear more than once (they appear in *bursts*) when they are relevant in a given document. We use the definition of burstiness from Church and Gale [CG95]:

$$B(w) = \frac{E_w(f)}{P_w(f \geqslant 1)}$$

where $E_w(f)$ is the mean number of occurrences of a given word w per document, and $P_w(f \geqslant 1)$ is the probability that w appears at least once in a document. The differences in frequency and burstiness are tested using the Mann-Whitney U test, which indicates if one population tends to have larger values than another. It is non-parametric, *i. e.*, it does not assume normality.

Differences in Frequency

Table 3.3 shows statistics related to word frequency in biography overviews for the LIWC categories. Note that, although the medians are very similar for each category, the U test compares differences in the distribution instead of differences in means or medians. If the test revealed significant differences, we calculated the *common language effect size* (ES) as the percentage of words that had a greater relative frequency for the dominant gender. Of the 20 categories under consideration, two of them (one top-level) shown significant differences between genders: *cogmech* (*cognitive processes*, ES = 63%) is dominated by men, while *sexual* (*sexual processes*, subcategory of *biological processes*, ES = 85%) is dominated by women.

Differences in Burstiness

Burstiness distributions in full biographies per semantic category are displayed in Table 3.4. There are three (two top-level) categories with significant differences, both dominated by men: *cogmech* (*cognitive processes*, ES = 60%), its subcategory *cause* (*causal processes*, ES = 71%), and *work* (*work concerns*, ES = 64%).

Overview of Results

In summary, in this section we found that words have similar meaning when referring to both genders, that there are qualitative differences in words most associated to them, and that a small number of the semantic categories show significant differences. Although this implies more similarities than differences in characterization of women and men, in the discussion section we elaborate over the importance of such differences and the implications of these findings.

3.5 NETWORK PROPERTIES

To study structural properties of biographies, we first built a directed network of biographies from the links between articles in the *Person* DBPedia class. This empirical network was compared with several null graphs that, by construction, preserve different known properties of the original network. This allows us to attribute observed structural differences between genders either to empirical fluctuations in such properties, such as the heterogeneous importance of historical figures, or to gender bias. To do so, we consider PageRank, a measure of node centrality based on network connectivity [BP98; For+07].

3.5.1 *Empirical Network and Null Models*

We study the properties of the directed network constructed from the links between 893,380 biographical articles in the *Person* class. After removing 192,674 singleton nodes, the resulting graph has 700,706 nodes and 4,153,978 edges. We use this graph to construct the following null models:

- *Random.* We shuffle the edges in the original network. For each edge (u,v), we select two random nodes (i,j) and replace (u,v) by (i,j). The resulting

Table 3.3: Word frequency in biography overviews. For each LIWC category we report vocabulary size, median frequencies, the result of a Mann-Whitney U test, and the three most frequent words. M and W mean men and women, respectively.

Category	V	Median (M)	Median (W)	U	Top-3 (M)	Top-3 (W)
social	498	0.04%	0.05%	-1.12	team (7.5%), son (7.0%), received (5.1%)	daughter (6.8%), received (5.9%), role (5.8%)
– family	43	0.03%	0.09%	-0.85	son (7.0%), father (5.0%), family (3.9%)	daughter (6.8%), family (4.7%), father (3.7%)
– friend	33	0.05%	0.05%	-0.58	fellow (2.0%), friend (0.8%), partner (0.8%)	fellow (1.8%), partner (1.1%), friend (0.8%)
– humans	59	0.13%	0.17%	-1.34	people (2.4%), man (2.2%), children (1.8%)	female (5.6%), women (5.3%), children (4.5%)
cogmech	1045	0.02%	0.02%	2.04*	became (10.8%), known (9.8%), made (8.1%)	known (10.3%), became (9.2%), since (8.1%)
– insight	354	0.02%	0.02%	0.73	became (10.8%), known (9.8%), become (2.2%)	known (10.3%), became (9.2%), become (2.0%)
– cause	182	0.02%	0.02%	1.31	made (8.1%), since (6.2%), based (3.3%)	since (8.1%), made (6.7%), based (4.2%)
– discrep	57	0.02%	0.02%	0.06	outstanding (0.4%), wanted (0.3%), besides (0.3%)	outstanding (0.5%), wanted (0.4%), hope (0.4%)
– tentat	151	0.01%	0.01%	0.85	appeared (3.2%), mainly (0.9%), mostly (0.8%)	appeared (6.8%), appearing (1.1%), mainly (0.8%)
– certain	110	0.03%	0.02%	0.92	law (2.7%), total (1.1%), completed (1.0%)	law (2.0%), ever (1.0%), completed (0.9%)
– inhib	229	0.01%	0.01%	1.75	held (4.2%), conservative (0.7%), control (0.5%)	held (3.1%), hold (0.6%), opposite (0.5%)
– incl	7	0.25%	0.29%	-0.06	addition (1.5%), open (0.8%), close (0.6%)	addition (1.7%), open (1.0%), close (0.4%)
– excl	6	0.11%	0.07%	0.48	except (0.2%), whether (0.2%), vs (0.1%)	except (0.2%), whether (0.1%), vs (0.1%)
bio	638	0.01%	0.01%	-1.63	life (3.9%), head (2.5%), living (1.1%)	life (4.7%), love (1.9%), living (1.7%)
– body	193	0.01%	0.01%	-0.60	head (2.5%), body (0.6%), face (0.5%)	head (1.5%), body (0.8%), face (0.6%)
– health	274	0.01%	0.01%	-0.40	life (3.9%), living (1.1%), hospital (0.9%)	life (4.7%), living (1.7%), health (1.2%)
– sexual	105	0.00%	0.01%	-3.02**	love (0.8%), passion (0.2%), gay (0.2%)	love (1.9%), sex (0.5%), lesbian (0.3%)
– ingest	122	0.01%	0.01%	-0.51	water (0.4%), food (0.3%), cook (0.2%)	food (0.5%), water (0.4%), cook (0.3%)
work	570	0.04%	0.03%	1.12	career (9.5%), team (7.5%), worked (6.4%)	career (8.1%), worked (6.6%), school (6.1%)
achieve	364	0.05%	0.04%	1.06	won (8.7%), team (7.5%), worked (6.4%)	won (13.0%), worked (6.6%), team (5.5%)

Table 3.4: Word burstiness in full biographies for LIWC categories. Columns are analog to Table 3.3.

Category	V	Median (M)	Median (W)	U	Top-3 (M)	Top-3 (W)
social	498	1.21	1.22	0.21	band (3.63), team (3.38), game (2.95)	team (3.40), women (3.14), role (2.96)
– family	43	1.31	1.35	-1.12	family (1.85), father (1.75), son (1.64)	family (2.02), mother (1.98), granny (1.87)
– friend	33	1.23	1.26	-1.06	friendly (1.86), buddy (1.66), guest (1.59)	guest (1.75), fellowship (1.54), buddy (1.53)
– humans	59	1.35	1.44	-1.00	sir (2.39), human (2.02), man (2.02)	women (3.14), mrs (2.33), lady (2.25)
cog-mech	1045	1.12	1.12	2.85**	open (2.37), law (2.36), decision (2.34)	open (3.28), revelator (2.75), law (2.31)
– insight	354	1.13	1.12	1.75	decision (2.34), logic (2.15), became (1.88)	revelator (2.75), became (1.86), ponder (1.86)
– cause	182	1.15	1.13	2.17*	force (2.05), made (1.92), production (1.85)	causation (2.29), outcome (2.04), production (1.82)
– discrep	57	1.10	1.14	-1.05	desir (1.49), outstanding (1.48), idealism (1.45)	outstanding (2.03), wanna (1.58), oughta (1.55)
– tentat	151	1.12	1.10	1.86	mysterium (1.96), puzzle (1.70), appeared (1.68)	appeared (2.02), bet (1.67), overall (1.63)
– certain	110	1.11	1.10	1.62	law (2.36), total (2.14), truth (1.50)	law (2.31), reality (1.55), total (1.52)
– inhib	229	1.10	1.10	1.09	fencing (2.28), security (1.92), defensive (1.89)	fencing (2.20), safe (2.16), blocker (2.02)
– incl	7	1.27	1.29	-0.45	open (2.37), inside (1.30), close (1.30)	open (3.28), close (1.31), additive (1.30)
– excl	6	1.27	1.20	0.48	vs (2.17), versus (1.32), whether (1.31)	vs (1.75), versus (1.35), whether (1.24)
bio	638	1.26	1.25	1.87	choke (5.18), lymphology (3.50), pelvimeter (3.00)	love (2.52), prostatic (2.50), hiv (2.33)
– body	193	1.27	1.26	1.24	pelvimeter (3.00), hip (2.15), pee (2.03)	prostatic (2.50), pelvimeter (2.00), tits (1.98)
– health	274	1.24	1.24	1.33	choke (5.18), lymphology (3.50), chiropractic (2.92)	hiv (2.33), choke (2.32), insulin (2.16)
– sexual	105	1.27	1.31	-0.51	gay (2.56), hiv (2.35), love (2.12)	love (2.52), prostatic (2.50), hiv (2.33)
– ingest	122	1.29	1.24	1.30	chew (2.39), cook (2.22), coke (2.11)	cookery (2.18), cooking (1.98), food (1.95)
work	570	1.23	1.20	2.62**	gre (4.54), team (3.38), dotcom (2.98)	pce (18.67), team (3.40), award (3.40)
achieve	364	1.15	1.15	0.54	team (3.38), win (3.19), king (2.64)	team (3.40), award (3.40), best (2.72)

network is a random graph with neither the heterogeneous degree distribution nor the clustered structure that the Wikipedia graph is known to have [Zla+06].

- *In-Degree Sequence.* We generate a graph that preserves the in-degree sequence (and therefore the heterogeneous in-degree distribution) of the original network by shuffling the sources of the edges. For each edge (u,v), we select a random node (i) and rewire (u,v) to (i,v). Each node has the same in-degree, or popularity, as the corresponding biography.
- *Out-Degree Sequence.* We generate a graph that preserves the out-degree sequence (and therefore the out-degree distribution) of the original network by shuffling the targets of the edges. For each edge (u,v) select a random node (j) and rewire (u,v) to (u,j).
- *Full Degree Sequence.* We generate a graph that preserves both in-degree and out-degree sequences (and therefore both distributions) by shuffling the structure in the original network. For a random pair of edges ((u,v), (i,j)) rewire to ((u,j), (i,v)). We repeat this shuffling as many times as there are edges. Note that although the in- and out-degree of each node is unchanged, the degree correlations and the clustering are lost.
- *Small World.* We generate a undirected small world graph using the model by Watts and Strogatz [WS98]. This model interpolates a random graph and a lattice in a way that preserves two properties of small world networks: average path length and clustering coefficient.

All null models have the same number of nodes $n = 700{,}706$ and approximately the same mean degree $k \approx 4$ as the empirical network. The Small World model has a parameter $\beta = 0.34$ representing the probability of rewiring each edge. Its value was set using the Brent root finding method in such a way as to recover the clustering coefficient of the original network.

Gender, Link Proportions and Self-Focus Ratio

For each graph, we estimated the proportion of links from gender to gender, and we tested those proportions against the expected proportions of men and women present in the dataset using a chi-square test. Table 3.5 shows the results. None of the null models show any bias in link proportions. The observed graph, on the other hand, shows a significant difference in the proportion of

Table 3.5: Comparison of edge proportions between genders in the empirical biography network and the null models. *M* and *W* mean men and women, respectively. All models share the same number of nodes, $n = 700{,}706$.

	Clust. Coef.	Edges	M to M	M to W	χ^2 (M to W)	W to M	W to W	χ^2 (W to W)	SFR
Observed	0.16	4,106,916	90.05%	9.95%	2.38	62.19%	37.81%	37.83***	6.55
Small World	0.16	2,775,372	84.45%	15.55%	0.00	84.15%	15.85%	0.01	5.41
Random	0	4,106,916	84.41%	15.59%	0.00	84.39%	15.61%	0.00	5.41
In Deg. Seq.	0	4,106,916	85.36%	14.64%	0.06	85.27%	14.73%	0.05	5.75
Out Deg. Seq.	0	4,106,916	84.43%	15.57%	0.00	84.37%	15.63%	0.00	5.42
Full Deg. Seq.	0	4,106,916	85.34%	14.66%	0.06	85.39%	14.61%	0.06	5.74

links from women biographies. In particular, articles about women tend to link to other women biographies more than expected ($\chi^2 = 40.54, p < 0.001$, Cohen's $w = 0.76$). Men biographies show a greater proportion of links to men and a lesser proportion to women than expected, but the difference is not statistically significant, although it has an impact on the estimated *Self-Focus Ratio* [HG09]. In our context, this ratio is defined as the relation between the sum of PageRank for men and the sum of PageRank for women. A SFR above 1 confirms the presence of self-focus, which, given the proportions of men and women in the dataset, is expected. In fact, given those proportions, the expected SFR is 5.41. Note that the null models have similar SFRs to the expected value, in contrast with the observed model with SFR of 6.55.

3.5.2 *Biography Importance*

As an approximation for historical importance in our biography network we considered the ranking of biographies based on their PageRank values.

Figure 3.6 displays the top-30 men and women according to their PageRank. Although the highest score entities present comparable scores, women present a faster decay than men. For instance, *Pope John Paul II*[9] (#10) has higher score

9 https://en.wikipedia.org/wiki/Pope_John_Paul_II

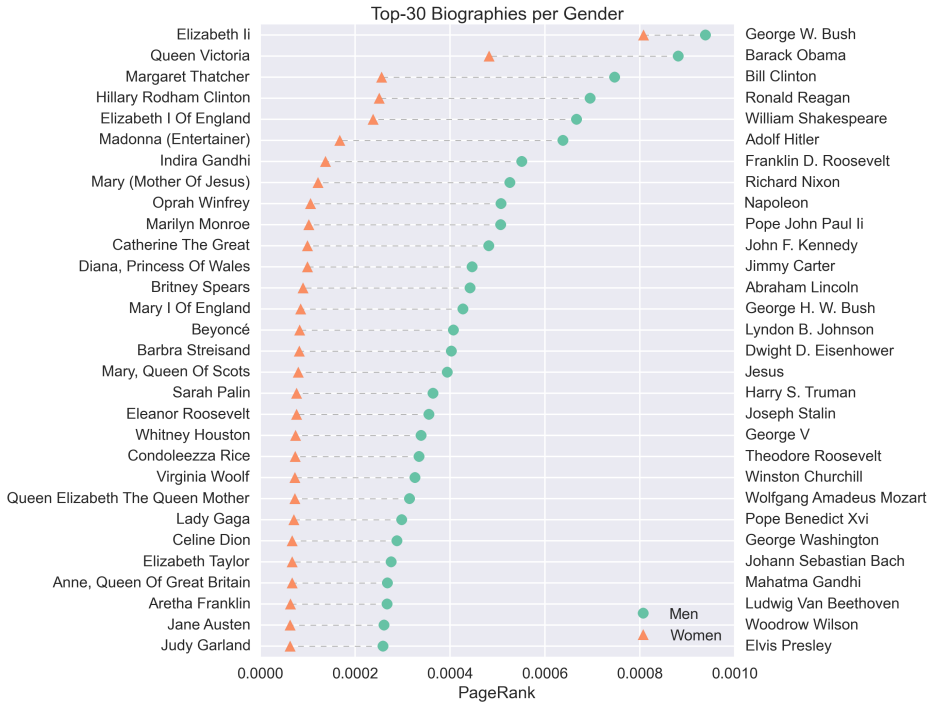


Figure 3.6: Top-30 biographies per gender according to PageRank.

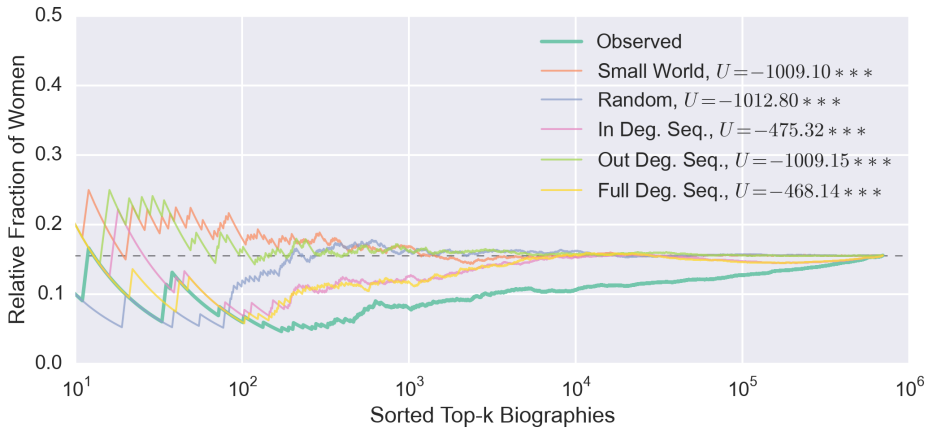


Figure 3.7: Women fraction in top biographies sorted by PageRank.

than *Queen Victoria*[10] (#2), and *Elvis Presley*[11] (#30) has higher score than *Hillary Rodham Clinton*[12] (#4). Our results are coherent with previous work: Aragón *et al.* [Ara+12] is more similar to ours because they consider PageRank only, while Skiena and Ward [SW14] considers other additional factors when ranking.

To compare the observed distribution of PageRank by gender to those of the null models, we analyzed the fraction of women biographies among the top-r articles by PageRank, for $r \in [10,\ 700, 706]$ (*i. e.*, we considered only nodes with edges). In the absence of any kinds of bias, whether endogenous to Wikipedia or exogenous, one would expect the fraction of women to be around 15% (the overall proportion of women biographies) irrespective of r. In the presence of correlations between popularity or historical importance and gender, we expect the ratio to fluctuate. But such fluctuations would also be observed in the null models.

The results are shown in Figure 3.7. While the null models stabilize around the expected value by $r \leqslant 10^4$, the proportion of women in the observed network reaches 15% only when the entire dataset is considered. This systematic under-representation of women among central biographies is not mirrored in the null models. We tested the differences between observed and null models using a Mann-Whitney U test, and found that the observed model is always significantly different (U values shown in Figure 3.7, $p < 0.001$ for all pairwise comparisons with the observed model, Holm-Sidak corrected). This implies a biased behavior that cannot be explained by any of the heterogeneities in the structure of the network preserved by the null models. For instance, even if men biographies tended to have more incoming links (as they do), or to be more densely clustered, those factors would not explain the lower centrality observed in women biographies.

10 https://en.wikipedia.org/wiki/Queen_Victoria
11 https://en.wikipedia.org/wiki/Elvis_Presley
12 https://en.wikipedia.org/wiki/Hillary_Rodham_Clinton

3.6 DISCUSSION

Even though we found more similarities than differences in characterization, in this section we contextualize those differences in social theory and history. We do this to understand why such differences exist, and whether they can be attributed to bias in Wikipedia or to a reflection of western society.

Meta-data

We found that there are statistically significant differences in biographies of men and women. Most of them can be explained because of the different areas to which men and women belong (mostly *sports* and *arts*, respectively), as well as the recency of women profiles available on Wikipedia. Other differences, like article length and article out-degree, although significant, have very small effect sizes, and depend on the person class being analyzed.

The greater frequency of the *spouse* attribute in women can be interpreted as specific gender roles attributed to women. A similar result on *Implicit Association* was obtained by Nosek, Banaji, and Greenwald [NBG02], as they found that Internet visitors tended to associate women to family and arts. Arguably, an alternative explanation is that people in the arts could be more likely to marry a notable spouse than people in sports. Yet, we found that the most common class was the generic one not assigned to any of those categories.

In terms of time, we found that the year 1943 marked a hit on the growth of women presence. According to Strauss and Howe [SH91], the post-war *Baby Boomers* generation started in 1943. The following generations are *Generation X* (1961–1981) and *Millenials* (1982–2004). The social and cultural changes embraced by people from those generations, plus the increased availability of secondary sources, might explain this growth. The growth started in dates nearby the French Revolution (1789–1799), where women had an important role, although they were oppressed after it [Abr75]. During these years seminal works about feminist philosophy and women's rights were published, like the works of *Mary Wollstonecraft* (1792) and *Olympe de Gouges* (1791). It is reasonable to assume that these historical events paved the way for women to become more notable.

Language

We found that the words most associated with men are mostly about sports, while the words most associated with women are to arts, gender and family. Of particular interest are two concepts strongly associated with women: *her husband* and *first woman*. These results are arguably indicative of systemic bias: the usage of *her husband* was found in concordance with our meta-data results and previous work by Bamman and Smith [BS14], and the already mentioned work on *Implicit Association* [NBG02]. These results can be contextualized in terms of *stereotyping theory* [PHK], as they categorize women, either as norm breaking (being the first is an exception to the norm) or as with predefined roles (being wives). As Fiske and Neuberg [FN90] indicate in their *continuum model of impression formation*, such categorization makes individuals more prone to stereotyping than those who are not categorized. The usage of *first woman* might indicate notability, but it also has been seen as an indicator of gender bias, as indicated by the Bechdel-inspired *Finkbeiner-test*[13] about scientific women, where it is explicitly mentioned that an article about a woman does not pass the test if it mentions *"How she's the 'first woman to ...'"* Despite being informal, the Finkbeiner-test raises awareness on how gender becomes more important than the actual achievements of a person.

To formalize the PMI analysis, we performed analysis based on semantic dictionaries of words. According to Nussbaum [Nus95], one possible indicator of *objectification* is the *"denial of subjectivity: the objectifier treats the object as something whose experience and feelings (if any) need not be taken into account."* This idea is supported as, in the overviews, men are more frequently described with words related to their *cognitive processes*, while women are more frequently described with words related to *sexuality*. In the full biography text, the *cognitive processes* and *work concerns* categories are more bursty in men biographies, meaning that those aspects of men's lives are more important than others at the individual level.

13 http://www.doublexscience.org/the-finkbeiner-test/

Presence and Centrality of Women

Women biographies tend to link more to other women than to men, a disproportion that might be related with women editing women biographies in Wikipedia, one of the reported interests of women editors [Sti13]. Since we are considering notable people, it is known that men and women's networks evolve differently through their careers [Jac89], not to mention the set of life-events that influence those changes like child-bearing and marriage (see a in-depth discussion by Smith-Lovin and McPherson [SM93]). Thus, link proportion between women cannot be attributed to bias in Wikipedia, as it seems to be more a reflection of what happens in the physical world.

We found that network structure is biased in a way that gives more importance to men than expected, by comparing the distribution of PageRank across genders. The articles with highest centrality, or historical importance [Ara+12], tend to be predominantly about men, beyond what one could expect from the structure of the network. As shown in Figure 3.7, there are women biographies with high centrality, but their presence is not a sign of an unbiased network: *"the successes of some few privileged women neither compensate for nor excuse the systematic degrading of the collective level; and the very fact that these successes are so rare and limited is proof of their unfavorable circumstances"* [De 12].

3.6.1 *Implications*

At this point, considering the *gender gap* that affects Wikipedia [HS13], it is pertinent to recall the concept of *feminine mystique* by Friedan [Fri10], developed from the analysis of women's magazines from the 50s in the United States, which were edited by men only. Fortunately, as discussed earlier, we have found women in different fields, mostly *arts*, in contrast to the *"Occupation: Housewife"* identified by Friedan [Fri10], as well as more similarities in characterization than differences. Moreover, the presence of women is increasing steadily and most of the differences found are not from an inherent bias in Wikipedia. Nevertheless, the identified language differences objectify women and the network structure diminishes their findability and centrality. Hence, the gender bias in Wikipedia is not just a matter of women participation in the community, because content and characterization of women is also affected. This is important,

for example, because Wikipedia is used as an educational tool [Kon10], and *"children learn which behaviors are appropriate to each sex by observing differences in the frequencies with which male and female models as groups perform various responses in given situations"* [PB79].

Editing Wikipedia and NPOV

Critics may rightly say that by relying on secondary sources, Wikipedia just reflects the biases found in them. However, editors are expected to write in their own words, *"while substantially retaining the meaning of the source material"*[14], and thus, the differences found in terms of language that objectify women are chosen explicitly by them. In this aspect, Wikipedia should provide tools that help editors to reduce sexism in language, for instance, by considering already existing manuals like [APA00]. Furthermore, their neutral point of view guidelines should be updated to explicitly include gender bias, because biased language is a clear violation of their guidelines.

Affirmative Action for Women in Notability Guidelines

The current notability guidelines for biographies in Wikipedia state: *"1. The person has received a well-known and significant award or honor, or has been nominated for one several times. 2. The person has made a widely recognized contribution that is part of the enduring historical record in his or her specific field."*[15] However, the boundary between not being notable according to sources and exclusion from history is blurred when evaluating the notability of women. For instance, consider a discussion about women in philosophy: *"Feminist historians of philosophy have argued that the historical record is incomplete because it omits women philosophers, and it is biased because it devalues any women philosophers it forgot to omit. In addition, feminist philosophers have argued that the philosophical tradition is conceptually flawed because of the way that its fundamental norms like reason and objectivity are gendered male"* [WS14]. Women, specially in historical contexts before 1943, should be targeted by affirmative actions that would allow them to appear in the content if they are not there, and be linked from other articles. We acknowledge that this is not easy, because

14 https://en.wikipedia.org/wiki/Wikipedia:No_original_research

15 https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)#Any_biography

relaxing notability guidelines can open the door to original research, which is not allowed. However, a correctly defined affirmative strategy would allow to grow the proportion of women in Wikipedia, make women easier to find, both through search (as it increases relevance) and exploratory browsing.

3.6.2 *Conclusions*

We studied gender bias in Wikipedia biographies. Our results indicate significant differences in meta-data, language, and network structure that can be attributed not only to the mirroring of the offline world, but also to gender bias endogenous to content generation in Wikipedia. Our contribution is a set of methodologies that detect and quantify gender bias with respect to content and structure, as well as a contextualization of the differences found in terms of feminist theory. As concluding remark, we discussed that Wikipedia may wish to consider revising its guidelines, both to account for the non-findability of women and to encourage a less biased use of language, which is a violation of its neutral point of view guideline.

Limitations

Our study has two main limitations. First, our focus is on the English Wikipedia, which is biased towards western cultures. However, a parallel work to ours by Wagner *et al.* [Wag+15] focused on hyperlingual quantitative analysis, and obtained similar results for other languages. Our methods can be applied in other contexts given the appropriate dictionaries with semantic categories, although our discussion remains to be applied, as it is culture-dependent. The second limitation is a binary gendered view, but we believe this is a first step towards analyzing the gender dimension in content from a wider perspective, given the social theory discussion we have made.

Future Work

At least three areas are ripe for further work. The first is the construction of editing tools for Wikipedia that would help editors detect bias in content, and suggest appropriate actions. The second is a study of individual differences among contributors, as our work analyzed user generated content without considering

who published and edited it. This aspect can be explored by analyzing how contributors discuss and edit content based on their gender and other individual factors. The last area is a further exploration of bias considering more fine-grained ontology classes and meta-data attributes. For instance, it may be possible that gender bias is stronger or weaker for different ontology classes (e.g., *Scientist* vs. *Artist*) or in biographies of people from different regions and religions. Finally it would be helpful to study whether gender bias depends on the quality of an article: does bias decrease with increasing number of edits or other measures of article maturity?

4

ENCOURAGING DIVERSITY AWARENESS

This part of the dissertation studies the effect of centralization (a systemic bias) on how people perceive geographically diverse timelines from micro-blogging platforms. In centralized countries, not only public policy, media and economic power are centralized, but people also turn their attention to central locations, impacting social media by biasing content. This in turn biases algorithms for content recommendation and search. To address this problem, we propose a methodology to evaluate centralization in micro-blogging platforms, and an information filtering algorithm to generate geographically diverse timelines. Through a case study with users in Chile, we confirmed that centralization from the physical world is reflected on its virtual population. Moreover, when evaluating the filtering algorithm, we found that the perception of diversity depends on whether users are in central locations or not, and that users from non-central locations do not see the diversity present on their timelines. We identify a diversity-awareness problem, which we propose to address using a mixture of diversity-balancing algorithms and diversity-encouraging user interfaces. We build an application for Chilean users of Twitter, whose interaction data is used to analyze their behavior from a central/non-central location perspective. We found that using information visualization techniques complements a diversity-balancing algorithm and that user perception is improved, as well as confirming that users behave differently.

4.1 INTRODUCTION

In his book on user experience, Bill Buxton said that *"in order to design a tool, we must make our best efforts to understand the larger social and physical context within which it is intended to function"* [Bux10]. In today's global Web, it is not clear if current social platforms consider those different contexts when building their user interfaces or defining their content-based algorithms. This would not be a problem in an uniform, unbiased world, but our world is neither uniform [HHM10] nor unbiased [MS96]. In fact, according to Gillespie and Robins [GR89], the technologies of communication that are supposed to shrink distances between communities are having the opposite effect, by constituting *"new and enhanced forms of inequality and uneven development"*. This is relevant in our context, as we work with Web platforms that are supposedly empowering users by removing physical barriers. In this part of this dissertation, we explore the effect of the systemic bias of *political centralization* through the following research question:

> *Does political centralization affect how people perceives information, and how people behaves when browsing informational content in microblogging platforms? If so, how to encourage geographically diverse exploration?*

Arguably, centralization is an organizational schema instead of a systemic bias. Given geographical and cultural contexts, centralization can be beneficial for the population and the economy, as discussed by Krugman [Kru99], but, while centralization is not inherently bad, *"over-centralization is often irreversible and hard to avoid"* [Kol13]. In some developing countries this organization tends to favor the most central locations by making public policy favor its needs [Bra99]. This is the case of Chile, a highly centralized country [GK08], where public policy, media and economic powers are centralized towards its capital region, *Región Metropolitana*. Although *RM* is indeed near the geographical center of continental Chile, it must be noted that the country spans over 4,300 km from north to south, with only 350 km from east to west at its widest point, in contrast with other centralized countries. This situation, with economical and geographical factors, makes Chile our focused country for study.

Through a focus on the geographical aspect of centralization, we defined a methodology that goes from the analysis of presence of centralization in micro-blogging platforms, to the definition of an information filtering algorithm that generates geographically diverse timelines. We evaluated this algorithm with users, and found that users from over-represented, central locations have a different sense of diversity than those from under-represented, peripheral locations. Given this difference, which made users from peripheral locations to consider our timelines (which were diverse by definition) as not diverse, we hypothesize that centralization generated a diversity-awareness problem. As a first step towards addressing this problem, we tested a known visualization of categorical news headlines by Weskamp [Wes04] in a novel and different context, *i. e.*, in diverse timelines with micro-posts. This interface makes diversity in timelines visually salient through the usage of *treemaps* [JS91], a well-known information visualization technique.

We evaluated this design in an exploratory application, named *"Aurora Twittera de Chile"*, a website where users were able to browse informative summaries of what was being discussed on Twitter in the entire country. An analysis of interaction data revealed that people engaged with it differently with respect to their geographical origin, and that our design encourages browsing of more diverse information from a geographical point of view. Our results demonstrate that political centralization is reflected from the physical world into micro-blogging platforms, and that it does affect the perception and behavior of users. In addition, our results support the claim that awareness of these social and physical contexts when designing algorithms and user interfaces improves user perception in the presence of these systemic biases.

4.2 BACKGROUND

The work presented in this part of the dissertation spans several research areas. We discuss these in relation to our aims, the positioning of our work, and approaches adopted.

Biases in Information Systems and Geography

Human behavior, in both on and off-line worlds, is affected by several biases, both cognitive and systemic. *Homophily* [MSC01] is the tendency to form ties with similar others, having similarity estimated from a variety of demographic factors (age, sex, location), topical interests or political leaning. This has been observed in tie-formation [HYG13] and topical influence [Wen+10] in micro-blogging plaftorms, and links between political parties in blog networks [AG05], among other contexts. Although in principle, personalized information systems present relevant content for users, the effect of cognitive biases like homophily is biasing those systems' output, which tend to provide mostly agreeable information, creating *filter bubbles* [Par11].

This paper deals with *political centralization* [Kol13], in particular considering its geographical aspect. Geography is an important attribute to be considered in the study of social networks and Web platforms, in particular Twitter. For instance, even though most of user ties are geographically local [QCC12], more than a third of mentions and links are inter-countries [Kul+12]. To understand how a virtual population is distributed, each user's location must be determined, a meta-attribute not always available. When geolocating users in micro-blogging platforms, the most basic approach is to query a gazetteer with the user's self-reported location [Mis+11; Hec+11], which usually comes in free-text form, and thus it is not normalized nor structured. More complex and accurate approaches involve entity recognition [Abe+12] and language models [CCL10; KMO11; AHS13], but they require a representative and often large corpus, which is not always available. Even though the lack of geographical diversity has not been perceived as a problem from a user-centered point of view (to the extent of our knowledge), it does lead to problems. In particular, in imbalanced and centralized contexts, location classifiers can become biased [Rou+13], and care must be taken when parameterizing machine learning algorithms.

The effects of centralization in the physical world have been studied with respect to public policy [Bra99], economy [GK08], among other aspects. Physical world phenomena and constraints reportedly affects virtual behavior and content. For instance, Takhteyev, Gruzd, and Wellman [TGW12] found that the number of international flights between countries is the best predictor of non-local ties in Twitter. In Twitter, centralization has been studied before at the

country level in the analysis of cultural differences, where countries with economic power are central in Twitter network connectivity [GMQ14; Pob+11].

Previous work by González-Bailón *et al.* [Gon+14] has found that sampled information sources from Twitter are biased in terms of network centrality, and that peripheral content is hard to find and under-represented in the sampled information stream. In our work, we evaluate if users from centralized locations behave differently than those from non-central locations in an application designed to mitigate the effects of centralization. We propose that, in a similar manner to filter bubbles [Par11], centralization makes content about central locations more prominent than it should compared with the population distribution. As noted by González-Bailón *et al.* [Gon+14], we take care into including content from non-central locations in order to avoid the bias already present in timelines sampled from the Twitter API.

Filtering and Representation of Diversity

There are several scenarios where searching and filtering micro-posts is needed, for example in situation tracking and crisis mapping [Mac+11], event monitoring [Dor+10; Mar+11] and faceted search [Abe+12]. In all these scenarios, relevance is considered as target attribute to maximize when filtering. Diversity is important, and can be ensured by minimizing similarity between items in a recommendation list [Zie+05], maximizing *information entropy* [Jos06] over a set of content features [DCC11], as well as context-specific diversification methods [MZR09]. Diversified sets have increased user satisfaction in book recommendation scenarios [Zie+05] but not in political scenarios unless users are *diversity-seekers*, usually a minority of users [MR10].

The recommendation and presentation of diverse and potentially challenging information has tried to mitigate bias effects [Far+10; MLR13; MR10; MZR09; Par+09], but the problems are far from solved, as users do not necessarily value diversity [MR10]. Most of this work was done in the context of political content. In a different area, Park *et al.* [Par+09] showed that the presentation of news headlines in clusters generates more clicks on news than non-clustered displays.

In this part of the dissertation we evaluate how users perceive diversity, informativeness and interestingness of timelines generated by an algorithm from previous work by Munson, Zhou, and Resnick [MZR09] and De Choudhury,

Counts, and Czerwinski [DCC11]. We focus on how user location influences those perceptions and how engagement differs in terms of location and user interface. We extend design guidelines by Park *et al.* [Par+09] by using treemaps [JS91] to maintain clustered representations of micro-posts while, at the same time, making diversity visually salient and noticeable. Treemaps have been used before to visualize content from micro-blogging platforms by Archambault *et al.* [Arc+11], although their approach is different to ours: they visualize clustered keywords, while we visualize entire tweets by using a design inspired by *Newsmap.jp* [Wes04], a treemap visualization of news headlines.

User Differences

The output of an algorithm can be diverse, but users do not necessarily perceive, understand or value this diversity [MR10]. To understand this, the study of *individual differences* [CCM00] is useful. In recommender systems, different users prefer different interaction methods depending on their own characteristics [KRW11]. Moreover, people from different cultures behave in different ways when communicating, not only on micro-blogs [GMQ14] but also on other forms of communication, like instant messaging [KFS06]. In our work, we propose that systemic biases introduce differences in how users behave, and thus, these differences should be accounted when designing systems in the same way as cultural differences.

4.3 FROM CENTRALIZATION TO INFORMATION FILTERING

We explain our methodology to balance diversity in information streams, such as the one offered by Twitter. Our methodology can be seen as a pipeline that starts with an analysis of a virtual population, builds a classifier to geographically annotate content, evaluates the classifier accounting for geographical diversity, and finally filters the content of streams to ensure geographical diversity. We focus on *timelines* from micro-blogging platforms. Although our definitions are general, we restrict ourselves to Twitter.

4.3.1 *Problem Definition*

We define our problem as follows:

1. Consider a set of tweets T_E related to an event E (defined as a set of hashtags and special keywords) relevant to a country C, with a set of locations L.
2. Given all users U who published tweets in T_E, predict (if possible) a location from L for all users $u \in U$.
3. Considering the users U_L who were geolocated, aggregate their interactions to find if geographical centralization is present.
4. Aggregate the content from U_L into location documents, and use these to build a location classifier P such that given an arbitrary tweet, predicts the *administrative location* related to its content.
5. Using the output from P applied to all tweets in T_E, filter T_E to produce a summary tweet set T_θ, with $|T_\theta| \leqslant |T_E|$, which is more geographically diverse; in other word such that $\text{geodiversity}(T_\theta) \geqslant \text{geodiversity}(T_E)$.

Geographical Diversity

We define geographical diversity as the normalized *Shannon entropy* [Jos06] with respect to a set of locations L (where $|L| > 1$):

$$\text{geodiversity} = \frac{-\sum_{i=1}^{|L|} p_i \ln p_i}{\ln |L|}$$

where p_i is the probability that a micro-post is related to a location ℓ_i. Geodiversity is 0 when all micro-posts are from one location only, and geo-diversity is 1 when all locations are represented equally.

Geolocating Users

To geolocate users, we rely on the self-reported location in user profiles. Instead of querying external services using profile locations as input [Mis+11], we build an ad-hoc gazetteer from official location names, lists of known toponyms extracted from Wikipedia [Hec+11; Rou+13] and labeled user profiles. Then, to geolocate a user u, we query the gazetteer with a normalized version of u's self-reported location. A normalized string (location) is defined as a lowercase

version of the original string, with redundant spaces and non-alphanumeric symbols removed.

4.3.2 *Interaction Graph and Centralization*

In [Kul+12], two locations become connected if someone from location A follows someone from location B. In our context this is not meaningful, because such connectivity may not convey an interaction between two users that is relevant to the event E [Wil+12]. Hence, we consider *1-way interactions* between locations through mentions and retweets [QCC12] by building an adjacency matrix:

$$M_{i,j} = \text{mentions}(\ell_i, \ell_j) + \text{retweets}(\ell_i, \ell_j)$$

where $\text{mentions}(\ell_i, \ell_j)$ is the number of tweets from location ℓ_i (those tweets whose author has been geolocated to that location) that mention one or more accounts from location ℓ_j, and $\text{retweets}(\ell_i, \ell_j)$ is the number of times that tweets from ℓ_j have been retweeted by users from ℓ_i.

From the adjacency matrix we build an undirected *interaction graph* with self-connecting edges removed, and estimate the edge weights with a normalized *geometric mean* of information flow:

$$w(i,j) = \frac{\sqrt{M_{i,j} \times M_{j,i}}}{\max\{\sqrt{M_{i',j'} \times M_{j',i'}} \mid \forall \ell_{i'}, \ell_{j'} \in L : i' \neq j'\}}$$

To estimate node importance in the interaction graph, we consider *random-walk weighted betweenness centrality* [New05]. We chose random walk instead of traditional betweenness centrality because information does not always follow geodesic paths, and weighted edges because those paths will have different importances based on the amount of pairwise interactions.

To analyze the existence of centralization as a system bias, we compare the observed centrality against a theoretical expected centrality in the interaction graph. We define the expected centrality as the *random-walk weighted betweenness centrality* [New05] in an population graph of locations, where each edge is

weighted according to the normalized geometric mean of each location's population:

$$w_{exp}(i,j) = \frac{\sqrt{pop_i \times pop_j}}{\max\{\sqrt{pop_{i'} \times pop_{j'}} \mid \forall \ell_{i'}, \ell_{j'} \in L : i \neq j\}}$$

Where pop_i is the physical population of location i. A considerable deviation in location centralities from the expectations is a strong signal of geographical centralization. To estimate such deviation, we estimate the C'_B measure by Freeman [Fre77], which considers the average of centrality differences between the most central node and all the others:

$$C'_B = \frac{\sum_{i=1}^{n} [C'_B(p^*_k) - C'_B(p^*_k)]}{n-1}$$

The values of C'_B vary between 0 (no centralization) to 1 (star-shaped network).

4.3.3 *Geographical Diversity and Classifying Tweets*

We assume that each location will have several local words and hashtags that characterize it, in addition to local event-specific keywords. These hashtags, among other words like place names, people names and vernacular words, will have more weight in their corresponding documents than global, non-local words. Hence, we consider that a tweet talks about a particular location if its content resembles or is similar enough to the aggregated content of that location.

To build a *location corpus* of $|L|$ *location documents*, we consider the set of geolocated users U_L. Each document is the aggregation of tweets originating from those locations, leaving out *replies*, *mentions* and *retweets* to avoid repeated content between different documents. We represent each location document d as a vector

$$\vec{d} = [w_0, w_1, \ldots, w_n]$$

where w_i represents the vocabulary word i weighted according to its locality by using TF-IDF [BR11a]:

$$w_i = \mathtt{freq}(w_i, d) \times \log_2 \frac{|L|}{|\ell \in L : w_i \in \ell|}$$

To predict a location for a given document $\vec{d}$, we build a feature vector $\vec{f}_d$ containing the similarity of $\vec{d}$ with each location document from the location corpus. In this way, we consolidate all similarities in a single vector:

$$\vec{f}_d = [f_0, f_1, \ldots, f_{|L|}]$$

where f_i is the cosine similarity between the document $\vec{d}$ and the location document $\vec{\ell}_i$:

$$\text{cosine\_similarity}(\vec{d}, \vec{\ell}_i) = \frac{\vec{d} \cdot \vec{\ell}_i}{\| \vec{d} \| \| \vec{\ell}_i \|}$$

We use the feature vectors and their corresponding author locations to train classifiers using *Support Vector Machines* (SVM) [CV95] and *Naive Bayes*.

A prediction is correct if the location predicted for a tweet t matches its author u's location, i.e, $t_\ell = u_\ell$. Although this approach may give *false positives* (a user tweets about other locations, for instance, media outlets publishing news) or *false negatives* (a user tweets about the event from a generic point of view), this assumption is also made in previous work [CCL10; Hec+11] because the usage of the self-reported location in geolocation allows to assign only one location to every user, which we find acceptable when considering events where users are expected to have a single location.

To find how different classifiers behave when considering geographical diversity, we consider p_i as the fraction of predictions for location i. To balance the classifier accuracy and geographical diversity, we define a *D-measure* as the harmonic mean between accuracy and geographical diversity:

$$D_\beta = (1 + \beta^2) \cdot \frac{\text{geodiversity} \cdot \text{accuracy}}{(\beta^2 \times \text{geodiversity}) + \text{accuracy}}$$

where β establishes the weight given to diversity: D_1 gives equal weight to accuracy and diversity, $D_{0.5}$ gives more weight to accuracy, and D_2 gives more weight to diversity. This flexible balance between diversity and accuracy is needed, because even though geographical diversity is important in our context, accuracy is needed; we do not want content to be erroneously classified. This approach was inspired by the *F-measure* that balances precision and recall in Information Retrieval [BR11a].

4.3.4 *Filtering Information Streams*

Given an event of interest E and a set of related tweets T_E, we generate a filtered tweet set T_θ, where T_θ contains s tweets. To maximize geographical diversity in T_θ, we consider a greedy algorithm from prior work by De Choudhury, Counts, and Czerwinski [DCC11]. This algorithm generates T_θ from an *information entropy* perspective, where entropy is estimated in terms of several features extracted from tweets. Since the complexity of those dimensions can be greater than those of geography (for instance, consider the number of hashtags in an event against the number of locations), the entropy contribution of these dimensions is higher than the entropy contribution of geography. Thus, diversity could still be near optimal levels even in the absence of geographical diversity.

Input Features

For each micropost t, we consider a vector representation $\vec{v}_t$ with the following features:

1. *Presence of links*: whether the micro-post contains a URL or not.
2. *Time bucket*: time passed since the start of E.
3. *Annotated hashtags*: topical information.
4. *Geography*: defined as the location its content is most likely to be about.
5. Author's number of *followers*.
6. Author's *hub dimension* ($\frac{\text{followers}}{\text{friends}}$).
7. Author's *global micro-post count*.
8. *Popularity*: number of times it has been republished by others.

All features are bucketed. We consider buckets based on the logarithms of each value when applicable.

Role of Popularity

As extension to the base algorithm [DCC11] we introduce popularity into the input features and into the tweet selection phase. In the original algorithm, tweet selection was random between all tweets that maximize entropy of the filtered set, while in our modification we rank candidate tweets by popularity. Popularity has been established as a valuable feature for tweet recommendation as it represents user interests [Che+10; Che+12]. However, given that popularity is

likely to link to population imbalance (*i. e.*, tweets from populated locations are likely to be more popular than those from less populated locations), we need to enforce geographical diversity.

Location Sidelines

Since the complexity of feature dimensions can be greater than those of geography (for instance, consider the number of hashtags in an event against the number of locations), the entropy contribution of these dimensions can be higher than the entropy contribution of geography. Thus, geographical diversity needs to be enforced by adding a *sidelining* step [MZR09] when considering tweets for selection, *i. e.*, tweets from a location previously selected in the previous iterations of the algorithm will not be considered during a given number of n turns.

Algorithm Definition

To maximize geographical diversity in T_θ, we consider a greedy algorithm: start by randomly selecting a micro-post t from the most popular bucket from T_E as initial seed in T_θ. Then, greedily increment T_θ by adding a selected micro-post from T_E. To select this micro-post, build a candidate set T_c where every micro-post t satisfies the following:

1. It has not been already added, i.e., it is not in T_θ.
2. Its addition to T_θ maximizes information entropy [Jos06].
3. Its location has not been considered before in at least n turns [MZR09].
4. It terms of popularity, it is on the most popular bucket.

Repeat until $|T_\theta| = s$.

The information entropy [Jos06] of a given tweet set T', with k different vector representations of its tweets ($k \leqslant |T'|$), is defined as:

$$H_{T'} = -\sum_{i=1}^{k} p(\vec{v}_{t_i}) \ln p(\vec{v}_{t_i})$$

This completes the description of each step of our proposed approach, which we now apply, as a case study, on the municipal elections held in Chile in 2012.

Algorithm 4.1 Geo. Diverse Information Filtering Algorithm.

```
INPUT: T ← set of micro-posts to be filtered
INPUT: s ← cardinality of resulting filtered set
INPUT: turns ← number of turns for sidelining
OUTPUT: T_θ ← filtered micro-post set
  FUNCTION GEODIVERSE_FILTERING(T, s, turns)
      T_θ ← list()
      sidelined ← dictionary()
    FOR ALL ℓ in L DO
          sidelined[ℓ] ← 0
    END FOR
      t ← random.choice(most_popular_microposts(T_E))
      T_θ.append(t)
      sidelined[t_ℓ] ← turns
    REPEAT
          T_c ← list()
    FOR ALL t in T_E not in T_θ DO
    IF max_entr(t, T_θ) and sidelined[t_ℓ] ⩽ 0 THEN
                  T_c.append(t)
    END IF
    END FOR
          t ← random.choice(popular_microposts(T_c))
          T_θ.append(t)
          sidelined[t_ℓ] ← turns + 1
    FOR ALL ℓ in L DO
              sidelined[ℓ] ← sidelined[ℓ] − 1
    END FOR
    UNTIL |T_θ| = s
    RETURN T_θ
  END FUNCTION
```

4.4 CASE STUDY: CENTRALIZATION IN CHILE

In this case study we apply the methodology defined in the previous section to a dataset of tweets from Chile, a country which suffers from political centralization [GK08]. In particular, centralization in Chile is characterized through geography at the *regional level*. Chilean regions are numbered (with roman numbers) from `I` to `XIV`, plus the capital region `RM` (*Región Metropolitana*, translated as *Metropolitan Region*), which is the most populated and centralized one. The administrative locations of Chile are defined according to the following hierarchy:

$$\text{municipality}\ (346) \rightarrow\ \text{province}\ (54) \rightarrow\ \text{region}\ (15)$$

For analysis we consider the 15 Chilean regions as basis. We refer to each of them as *region* or *location* equally.

4.4.1 *Dataset: Municipal Elections in Chile*

The dataset is composed of tweets crawled on October 28th, 2012, in the context of municipal elections held in Chile that day. The event had a distinctive hashtag (*#municipales2012*), which, among other related hashtags (e.g. *#tudecides*), keywords (e.g. *vote*), location and candidate names, were used as queries for the *Twitter Streaming API*.[1] Example terms used as queries are displayed on Table 4.1. This dataset being about local elections happening nation-wide makes it ideal to study effects of centralization.

In total, we have 157,648 users, who published 724,890 tweets. Table 4.2 summarizes the characteristics of those tweets. The most frequent terms in the dataset are shown in Figure 4.1. Query terms are colored green, while other terms are colored grey. Frequency was estimated ignoring word repetitions in each tweet. As expected, the most frequent term is *#municipales2012*. The figure shows several kinds of terms, such as keywords (*e. g.*, *democracia*, *comuna*, *resultados*), people (*e. g.*, journalists *@tv_mauricio*, *@patricionavia*, candidates *@josefaerrazuriz*, *sabat*, *labbe*, *zalaquett*), hashtags (e.g., *#tudecides*, *#labbe*) and

1 https://dev.twitter.com/docs/streaming-apis

Table 4.1: Example terms used as queries.

Type	Query Terms
Hashtags	#municipales2012, #túdecides, #yovote
Vocabulary	elecciones municipales, elección, abstención, vocal de mesa, mesa, urna, deber cívico, alcade, alcadesa, alcaldía, concejal, voté, voten, votamos ...(tenses of *to vote*)
Politics	concertación, alianza por chile, servel
Politicians	labbe, errazuriz, zalaquett, tohá, sabat, ...
Locations	chile, santiago, concepción, valparaíso, ...

Figure 4.1: *Wordcloud* of frequent keywords.

place names, including peripheral ones (*e. g.*, *santiago*, *iquique*, *#valdiviacl*) to reduce bias in API results [Gon+14].

Using a list of 2011 toponyms from all locations in Chile, we were able to geolocate 33.67% of users at regional granularity using their self-reported location. Those users published 43.27% of the crawled tweets. Even though in 2012 RM held 40.5% of the population [Nat14], in our dataset it holds 56.6% of user accounts, and this difference in proportions is significant according to a chi-square test ($\chi^2 = 11.08$, $p < 0.001$), meaning that the Chilean population in Twitter is more imbalanced than in the physical world.

Table 4.2: Main types of data crawled during the `#municipales2012` event.

Data	#	% of Total	% Geolocated Region
Users	157,648	100.00	33.67
All Tweets	724,890	100.00	43.27
ReTweets	253,582	34.98	46.46
With Mentions	192,828	26.60	43.51
With Hashtags	227,868	31.43	48.03
With Links	76,440	10.55	43.40
With Lat.,Lon.	50,489	6.97	50.08
With Lat.,Lon. or Hashtags	266,106	36.71	48.14

From Table 4.2 note that the fraction of tweets with geographical coordinates is small (6.97%) and the fraction of tweets with hashtags is less than a third of the entire information space (31.43%). Both attributes combined comprise 36.71% of the dataset. This means that if an information seeker has a query related to a specific place, a query using hashtags or the native geolocation of the platform would allow her/him to find at most 36.71% of tweets.

Dataset Cleaning

Given that elections are fairly general, *i. e.*, not specific to politics (*e. g.*, musical shows where the public votes for their favorite musician), the dataset contained some noise, *i. e.*, tweets from other countries and in different languages (*e. g.*, *vote* is also a valid english word). To clean the dataset, we manually inspected crawled tweets to find common unrelated keywords (*e. g.*, `bieber`, `lovato`, `romney`, `obama`, `#xfactor3`, and so on) and unrelated locations in user profiles (*e. g.*, *Argentina*, *Perú*, *España*, *Spain*, and so on). We removed all tweets that contained at least one of the blacklisted keywords, as well as users from those blacklisted locations. In addition, we performed language detection using *n-grams* [LB12], and removed all tweets which were not determined to be in Spanish. We also removed tweets that were determined to be *check-ins* into location social networks like *FourSquare*, even if they were made in Chilean territory, as not all check-ins were related to the event.

User Connectivity and Time of Registration

In Figure 4.2, we explore the properties of all users, regardless of their geolocation. On the left, we consider the complimentary cumulative distribution functions (CCDF) of the number of followers and friends according to accounts from `RM` and `NOT-RM`, showcasing similar distributions to those of [Kwa+10]. The estimated power-law [ABP14] PDF parameters are `RM` ($\alpha_{\text{follow}} = 1.20$ and $\alpha_{\text{friends}} = 1.18$) and `NOT-RM` ($\alpha_{\text{follow}} = 1.21$ and $\alpha_{\text{friends}} = 1.18$). On the right, we show the distribution of registration date of geolocated accounts. We see accounts from the beginning of Twitter until the day of the studied event. The biggest peak in terms of account creation corresponds to the week of February 27th, 2010, when an earthquake affected Chile. Twitter played a major role in information diffusion in the following days [MPC10].

4.4.2 *Virtual Population, Centralization and Content*

In this section we consider the geolocated tweets at regional level in terms of the author location, i.e., the 33.67% of users who contributed 43.27% of the event tweets. Figure 4.3 shows the population distribution at the top row of charts. On the left, the virtual population per region is showcased through the absolute number of user accounts found at each location; on the right, the *user rate* (relative number of accounts per 1,000 inhabitants is shown). It can be seen that, while the location populations differ in orders of magnitude, the user rates do not.

To explore the representativity of the sample, we estimate *Pearson product-moment correlation coefficients* between populations and rates, a measure of the linear dependence between two variables. Its value is in the range $[-1, 1]$, where -1 implies total negative correlation, 0 implies no correlation, and 1 implies total positive correlation. The estimated correlations are:

1. $r_{\text{pop}} = 0.95$ ($p < 0.001$), the correlation between the virtual and physical population from 2012 [Nat14]. Because of population imbalance in both worlds (physical and virtual), we consider the logarithms of population counts. See Figure 4.3 Bottom Left.

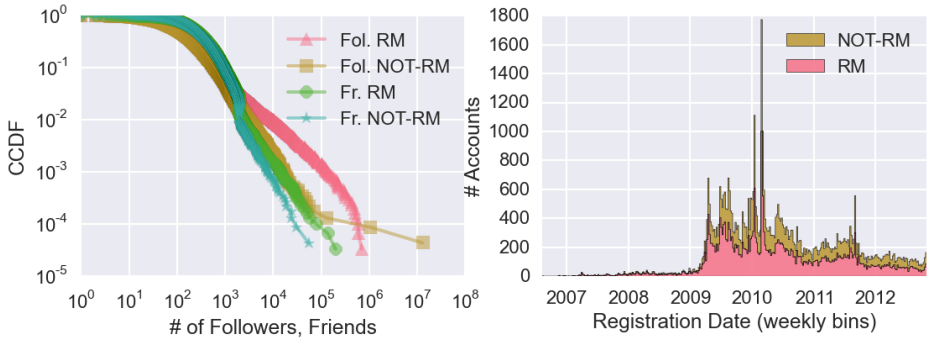


Figure 4.2: User connectivity and time of registration.

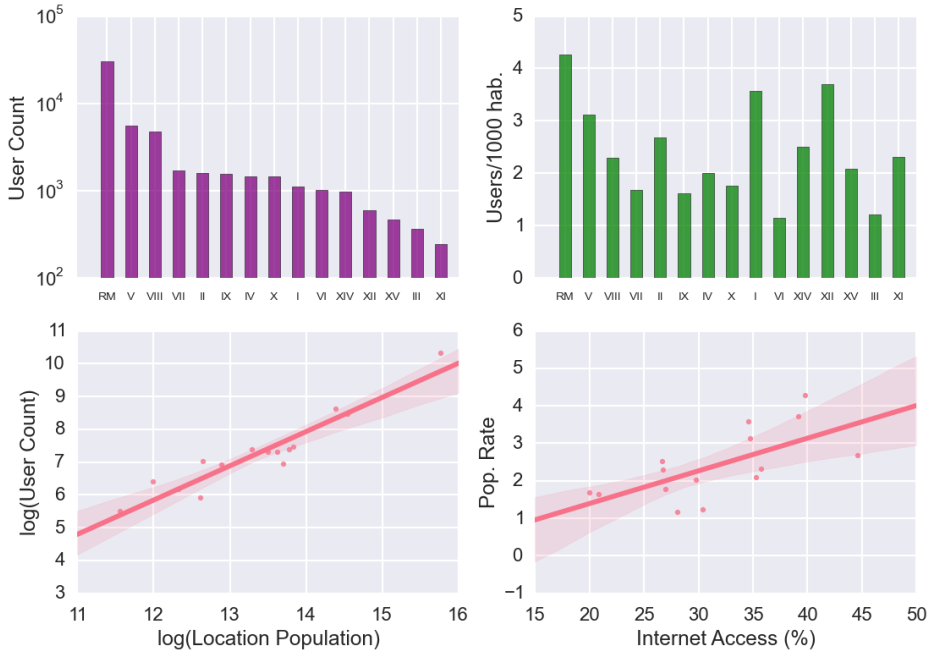


Figure 4.3: Top: Distributions of population according to Chilean regions (left) and user rates per 1,000 inhabitants (right). Bottom: linear regressions of logarithms of physical population [Nat14] with Twitter accounts (left), Internet access rate [Min11] with Twitter account rate (right).

2. $r_{rates} = 0.66$, ($p < 0.01$), the correlation between the *user rate* virtual and the *household Internet Access Rate* in Chile [Min11]. See Figure Figure 4.3 Bottom Right.

Therefore, our sample is *spatially representative* of the physical population at the regional level. This representativeness means that the amount of tweets considered is sufficient for the content analysis performed in the next sections.

Adjacency Matrix and Interaction Graph

We built and adjacency matrix of 1-way interactions and its corresponding graph. Figure 4.4 displays the adjacency matrix as a flow diagram between regions. Each region appears twice, once as a source, and once as a target. Each edge encodes $M_{i,j}$ from the adjacency matrix. An edge color encodes the target region: *green* encodes interactions when a region interacts with itself ($i = j$), *brown* encodes interactions when the target region is *RM* ($l_j = RM$), and *gray* encodes all remaining interactions. The median proportion of tweets emitted to other locations is 57.44%, while the median proportion of tweets received from other locations is 39.23%. We see that *RM* emits a majority of interactions (58.24%), and that the amount of interactions where *RM* is target is higher (71.48%). The average proportion of tweets emitted at *RM* is 49.47%, and the average proportion of tweets received from RM is 26.16%. *RM* is the only location whose relation between emitted to other locations and received from other locations is less than 1 (0.33), while the median is 2.40 and the max is 2.95. This behavior hints the possibility of centralization.

To confirm centralized behavior, we estimated *random walk betweenness centrality* [New05] on the location interaction graph built from the adjacency matrix. As defined in our methodology, we considered expected and observed weights on the graph edges, based on physical population and interactions, respectively. Figure 4.5 displays the differences found in centrality found for each location. Indeed, *RM* is the most central location (0.76 observed versus 0.19 expected), having the only increase and highest absolute difference between observed and expected centralities. Moreover, the observed centralities are statistically different to the expected centralities according to a Mann-Whitney U test ($U = -3.05$, $p < 0.01$). Even when considering population imbalance, the expected interaction graph is not centralized ($C'_{B(EXP)} = 0.13$), whereas the

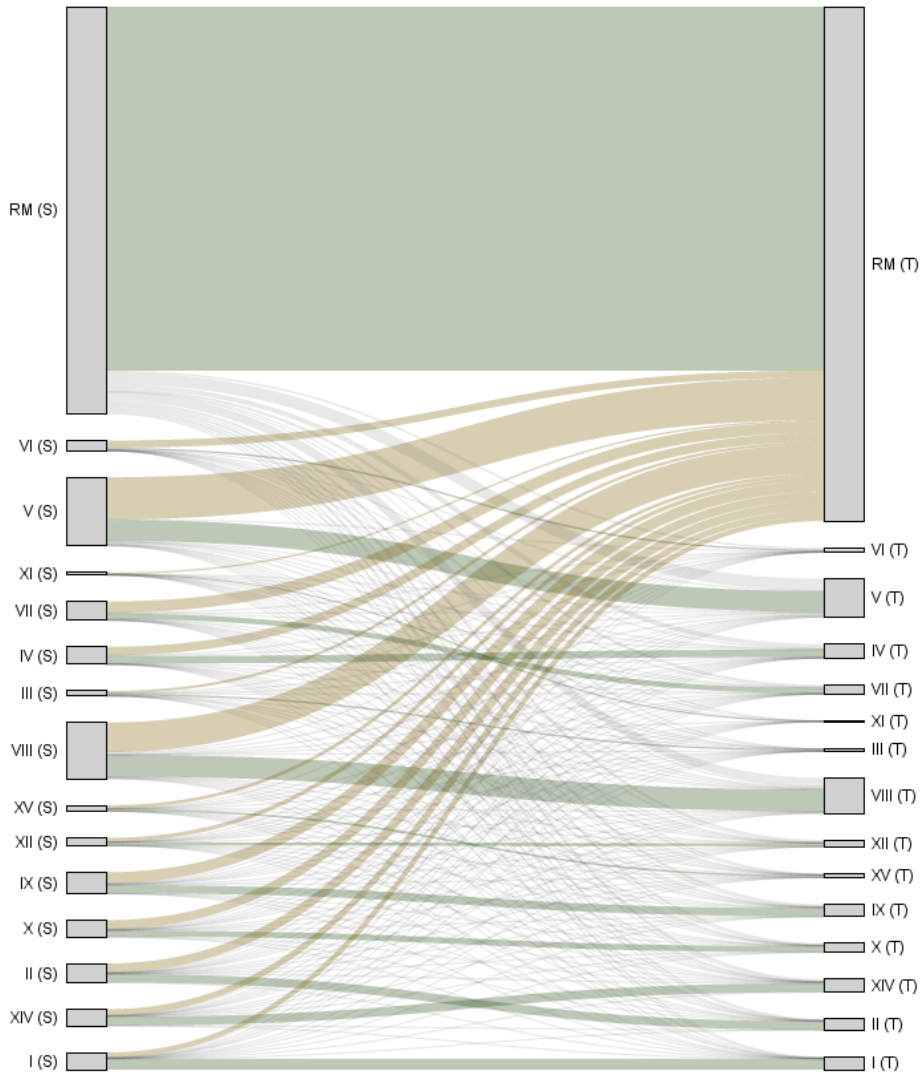


Figure 4.4: Adjacency matrix between locations represented as a flow diagram. On the left, each location node is a source of interaction, while on the right each location node is a target of interaction. Thus, each location appears twice: when emitting tweets, and when receiving tweets. The size (height) of each node is proportional to the total amount of interactions emitted/received. Edge color encodes target location: green encodes interaction with itself, brown encodes interaction with RM, gray encodes all other interactions.

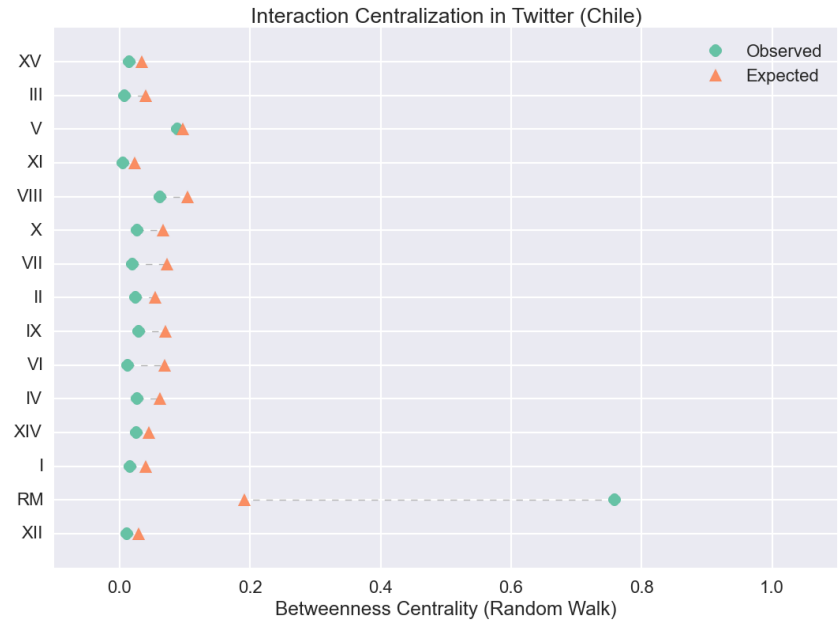


Figure 4.5: Differences in expected and observed centralities estimated on the interaction graph.

observed interaction graph is highly centralized ($C'_{B(OBS)} = 0.73$) Therefore, there is a clear centralization in location interactions between Chilean regions.

Regional Content

For each region we created a *location document* that concatenated all tweets published by authors geolocated in it. To avoid noise, we discarded keywords that appeared in less than five different tweets, and we considered each keyword once per tweet. Table 4.3 displays the most frequent and the most discriminative terms according to TF-IDF.

Most of the found discriminative keywords can be categorized in:

- Toponyms: *#laserena* (IV), *coyhaique* (XI), etc.
- Candidate names: *#soria* (I), *arellano* (VI), etc.
- Adaptations of event hashtags: *#municipalesmag* (XII), *#municipalesfm* (V), etc.

Table 4.3: Top-5 Frequent and Discriminating Keywords per Region.

Region	Popular Keywords (TF)	Discriminating Keywords (TF-IDF)
I	#iquique #municipales2012 soria iquique votar	#iquique #soria tarapaca @hombrederadio myrta
II	#municipales2012 votar #antofagasta votos voto	#antofagasta hernando @antofagastatv @karenrojov #karenrojo
III	#municipales2012 votar #copiapo votos copiapo	#copiapo cicardini #atacama maglio copiapo
IV	#municipales2012 votar votos coquimbo mesa	@elobservatodo #laserena @eldia_cl #coquimbo #ovalle
V	#municipales2012 votar votos labbe mesa	sumonte urenda #municipalesfm #valparaiso #quillota
VI	#municipales2012 votar votos labbe chile	rancagua #rancagua arellano @alcaldesoto llancao
VII	#municipales2012 votar votos talca voto	#talca talca #linares #curico #talcavota
VIII	#municipales2012 votar votos voto labbe	@cristian_quiroz zarzar #chillan hualpen @armstrongconce
IX	#municipales2012 votar #temuco votos mesa	#araucania #temuco #araucaniaelige huenchumilla @obduliovaldeben
X	#municipales2012 votar votos puerto mesa	#puertomontt #osorno @gervoyparedesr #puertovaras #todoesposible
XI	votar #municipales2012 coyhaique votos chile	huala coyhaique #coyhaique #aysen acevedo
XII	votar #municipales2012 #puq arenas punta	#puq #puqvota #municipalesmag #polartv @emilioboccazzi
RM	#municipales2012 votar labbe votos providencia	#melipilla quilicura #actualidad #puentealto pudahuel
XIV	#valdiviacl #municipales2012 votos sabat valdivia	#valdiviacl #lu2012 #uach @carlosamtmann n_n
XV	#municipales2012 #arica arica votar #aricavota	#aricavota #arica valcarce rocafull alamo

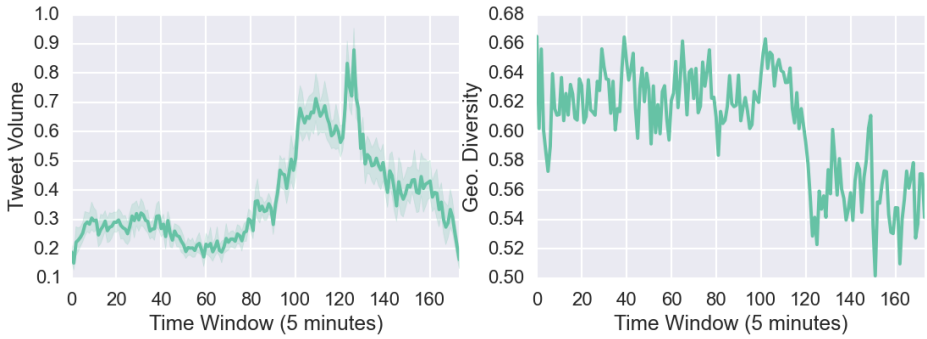


Figure 4.6: Time-series of normalized tweet volume from regions through the event (left) and geographical diversity of those tweets (right).

- Media accounts and hashtags: *@antofagastatv* (II), *#polartv* (XII), etc.

The existence and identification of local keywords validate our assumptions regarding how we defined our vocabulary.

Tweet Volume and Geographical Diversity

To explore temporal behavior and geographical diversity of the sample, Figure 4.6 shows regional tweet volume and dataset geographical diversity. During the day activity increased steadily as elections were being held in a similar way in every region, meaning that the event had a common structure in the whole country. At night, specific events regarding unexpected and controversial election results raised the level of activity above expectation. Figure 4.6 Left displays how the tweet volume varied during the event: as elections progressed during the day, tweet volume slowly increased, and in the afternoon, when results started to emerge, the tweet volume reached its peak. Geographical diversity was always in the range $[0.50,\ 0.67]$ (see Figure 4.6 Right). For comparison, note that the geographical diversity of the 2012 Chilean population is 0.77, which means that, although there is geographical diversity in the dataset, it is below the value one would expect given the population distribution. The observed decay in diversity is explained by the unexpected and sudden defeat of several candidates in some locations, shifting the discussion in a natural way towards fewer locations. The existence of geographical diversity means that it is possible to find

content related to all locations, making feasible the application of a filtering algorithm to generate geographically diverse timelines.

Latent Topical Diversity

We have seen that there is geographical diversity in our sample, as well as the existence of locally relevant keywords and hashtags. However, it is not clear if geographical diversity is related to popularity, one of the dimensions that is considered by our information filtering algorithm. To explore this, we built a topical space by considering *micro-blogs* or *user documents* (e.g. the concatenation of tweets by a user) for topic modeling using *Latent Dirichlet Allocation* [BNJ03], as in previous work [RDL10]. LDA is a generative model that explains word usage in a document by contributions of several latent topics, allowing us to estimate the probability that a topic T contributes words to a given document L, *i. e.*, $P(T \mid L)$.

To estimate geographical diversity of latent topics, we need the probability that a location L_i contributes to a topic T. To estimate $P(L_i \mid T)$ we use Bayes' Theorem and the law of total probability:

$$P(L_i \mid T) = \frac{P(T \mid L_i)P(L_i)}{P(T)} = \frac{P(T \mid L_i)P(L_i)}{\sum_j P(T \mid L_j)P(L_j)}$$

where $P(T \mid L_i)$ is estimated from the LDA model and $P(L_i)$ is the probability that a tweet comes from location L_i.

We estimated LDA with $k = 200$ latent topics over the set of user documents with the *gensim* software library [ŘS10]. Only a fraction of latent topics are geographically diverse, as only 59 topics contribute to at least one location. Topics that do not contribute to any location document can be *junk topics* [AlS+09] (recall that topics that contribute to only one location have geographical diversity 0 but still contribute). The mean number of locations a topic contributes to is 3.25, and the mean geographical diversity is 0.14, which is very low, meaning that most topics are about local discussion, although other topics are highly diverse, having a geographical diversity as high as 0.89. This means that it is possible to have geographical diversity while still maintaining topical diversity, as a portion of latent topics contribute to more than one location.

An analysis of how users are interested in those 59 topics is needed. We estimated the *Spearman rank coefficient* ρ between retweets and geographical di-

Table 4.4: Top-15 Geographically Diverse Latent Topics with Top-5 Contributing Hashtags and Mentions.

Diversity	Locations	ReTweets	Contributing Hashtags	Contributing Mentions
0.89	14	7480	#municipales2012, #copiapo, #linares, #tierraamarilla, #cerrillos	@kathybodis, @rau_2012_, @ipoduje, @cerrilloscl, @cooperativa
0.78	14	5444	#municipales2012, #macul, #concon, #municipalesterra2012, #municipalesconcon	@biobio, @carabdechile, @reddeemergencia, @rurendah, @tv_mauricio
0.69	8	4603	#municipales2012, #puertovaras, #vuelenalto, #quemchi, #bahamonde	@soychilecl, @radiousach, @aberger2012, @marcelomorane, @jantoniobotto
0.60	15	4883	#vuelaaltolabbe, #municipales2012, #yovote, #nodalomismo, #corta	@ppiippee, @lecarini, @copan, @dexter_25, @jhendelyn
0.60	12	6832	#fail, #municipales2012, #twd, #iiiiiii, #centralismo	@tjholt, @alvarez_monse, @informadorchile, @arenita_eventos, @alegriagonzaa
0.55	12	10320	#municipales2012, #yonoprestoelvoto, #tiempodetomas, #cuentaconmigo, #udechile	@mariana_gat, @_eloisagonzalez, @citizenalmeida, @ucvtvnoticias, @s_schwartzmann
0.54	13	2530	#municipales2012, #tudecides, #vuelaaltolabbe, #yovote, #chv	@randudog, @televisivamente, @arayabirkner, @condonito_chile, @cooperativa
0.50	11	8520	#municipales2012, #leyseca, #salvadorallende, #votovoluntario, #seguro	@sebastianpinera, @corphumanas, @chucknews, @opasten, @comunidadmujer
0.48	4	5700	#estaquearde, #municipales2012, #sisepuede, #yovote, #mujerazo	@youtube, @rafa_cavada, @katomagna, @ariel_levy, @charrisonh
0.44	5	6075	#municipales2012, #vamos, #vota, #tolerancia, #tudecides	@bdelamaza, @rorritz, @don_absurdo, @mhilsenrad, @macahederra

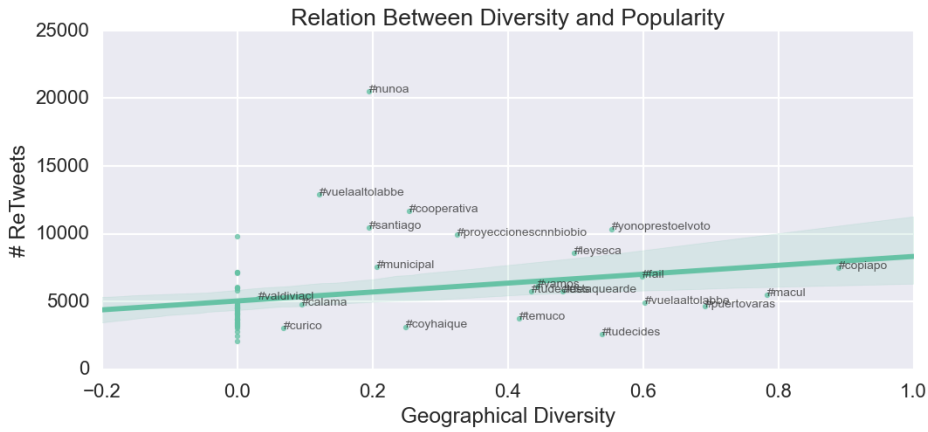


Figure 4.7: Relationship between geographical diversity and retweets. Each dot is a latent topic. Latent topics that contribute to more than one location are labeled with a highly contributing hashtag according to LDA.

versity of each topic. The Spearman rank coefficient is a measure of how monotonically related two variables are, and its value ranges between -1 and 1, with 0 implying no relationship. The number of retweets and geographical diversity show a high *Spearman rank coefficient* of $\rho = 0.42$ (with $p < 0.001$). Hence, there is a moderate positive monotonic relation between geographical diversity and popularity. This relation is visualized in Figure 4.7. Note that the most popular topics are highly diverse, although they are still less diverse than the expected diversity of 0.77 according to population distribution. In the figure, those topics contributing to at least two locations are labeled with their most or second most prominent hashtag (we display the second one if the first is the event hashtag, *#municipales2012*).

Table 4.4 contains the top-15 geographically diverse latent topics, as well as the five most contributing hashtags and mentions per topic. As expected, one of the most common contributing hashtags is *#municipales2012*. Some remarks about the contributing hashtags and mentions:

- In the most diverse topic, the other four contributing hashtags are location names (*e. g.*, *#copiapo* and *#tierraamarilla*), and in the second most

diverse topic, two are location hashtags (*#macul* and *#concon*) while another is a local variation of the event hashtag (*#municipalesconcon*).
- General, commentary hashtags appear in many locations (*e. g.*, *#fail*, *#estaqueard*e, *#nodalomismo*, *#yovote*, and *#yonoprestoelvoto* appear in 13 locations; *#yovoto* appears in 12 locations).
- Controversial candidates and related accounts/hashtags appear in many topics: *#vuelaaltolabbe* is related to *Cristián Labbé*, *#iiiiiii* is related to *Pablo Zalaquett*. Both candidates are based in Santiago, but their electoral defeats where discussed in the whole country.
- National level media outlets (*e. g.*, *@cooperativa*, *@biobio*, *@soychilecl*) appear on the most diverse topics.
- Many non-political, but specialized accounts appear through all the topics (*e. g.*, *@ipoduje* is an urbanist, *@reddeemergencia* is a network of emergency situations, *@jhendelyn* is a television model, *@alvarez_monse* is a journalist, and so on).

This means that highly diverse latent topics have the potential to be involved in multi-regional discussion, while at the same time being popular. This discussion can be political (according to the topic of the main event), but also it may involve people and topics which are not inherently political, and that serve as connectors between locations. In this aspect, it makes sense to present geographically diverse content to users, as we have found that diversity is moderately related to popularity. Whether this popularity is a consequence of geographical diversity or not, is beyond the scope of this part of the dissertation.

4.4.3 *Classifying Tweets Into Locations*

Despite the spatial representativity of our sample, the usage of a gazetteer to geolocate users leaves out a considerable amount of tweets. To be able to capture the potential content present in those tweets, there is a need to classify them into locations, to be considered for selection in our filtering algorithm. Following our methodology, we used the location corpus built previously to create and evaluate location classifiers using as input the feature vectors associated to its documents.

Evaluation of Location Classifiers

We evaluated the following classifiers using a 10-fold stratified cross-validation: SVM Linear Kernel (with a *one versus one* multi-class strategy), SVM Linear Kernel (with a *one versus all* multi-class strategy [RK04]), SVM RBF Kernel and Naive Bayes. All classifiers are from the *scikit-learn* library [Ped+11] and used with their default configurations. We divided the set of tweets from geolocated users in 10 groups, maintaining the proportions of locations' tweets in each group, and then ran 10 iterations to evaluate the classifiers. In each iteration we trained each classifier using 9 groups and tested predictions with the remaining group. In this way, each tweet was used nine times for training and one time for evaluation. We did not consider retweets and replies to avoid duplicate tweets in training and evaluation sets. Then, we estimated the geographical diversity of the set of predictions of each classifier, and calculated the *D-measure* at $\beta = \{0.5, 1, 2\}$. To evaluate our approach, we considered the following baselines:

1. *Trivial Classifier*, which predicts the most common location in the dataset (*RM*).
2. *Best Cosine Similarity*, which predicts the location with the highest cosine similarity between the tweet content and the location documents, as in [GP13].
3. *SVM* and *Naive Bayes* classifiers trained with *bag of words*, with vocabulary size $51,354$.

Evaluation Results

The results are reported in Table 4.5. In terms of accuracy, the SVM-based classifiers had the best performance, which aligns with previous work in imbalanced populations [Rou+13]. However, not all classifiers shown diversity: some of them have null entropy, which means that they are behaving in the same way as the trivial classifier, as depicted in Figure 4.8 (*Naive Bayes*) and Figure 4.9 (*Naive Bayes*, *SVM Linear1vs1* and *SVM RBF*). To analyze the trade-off between accuracy and diversity, we consider $D_{0.5}$, in which the best scores are for *Cosine Similarity* (0.60) and *SVM Linear1vsAll* (0.57). Which one of the two is better will depend on the situation. SVM has proven to be robust at different geographical granularities in the presence of imbalance [Rou+13], while *Cosine Similarity* has not. In fact, in this dataset, a preliminary experiment shown that increasing the

Table 4.5: Evaluation results at regional level of our classifiers using a 10-fold stratified cross validation. Classifiers prefixed with *BoW-* use normalized *bags of words*, whereas the other classifiers use TF-IDF weighting according to locations.

Approach	Accuracy	Geo. Div.	$D_{0.5}$	D_1	D_2
SVM Linear1vsAll	0.67	0.36	0.57	0.47	0.39
Naive Bayes	0.59	0.00	0.00	0.00	0.00
SVM RBF	0.68	0.31	0.55	0.43	0.35
SVM Linear1vs1	0.68	0.31	0.55	0.43	0.35
Trivial	0.59	0.00	0.00	0.00	0.00
Cosine Similarity	0.60	0.55	0.59	0.58	0.56
BoW-SVM Linear1vs1	0.59	0.00	0.00	0.00	0.00
BoW-SVM Linear1vsAll	0.67	0.26	0.51	0.38	0.30
BoW-SVM RBF	0.59	0.00	0.00	0.00	0.00
BoW-Naive Bayes	0.59	0.00	0.00	0.00	0.00

geographical granularity decreased its accuracy [GP13]. Additionally, note that even without our content features, the *SVM Linear1vsAll* classifier had a considerable amount of diversity, but still lesser than with our approach ($D = 0.51$).

4.4.4 *Overview of Analysis*

In this section, we focused on a case study, the Chilean municipal elections held in 2012. We found that centralization is reflected in Twitter when the virtual population is centralized. We also found that geographical diversity exists, that geographically diverse content is moderately correlated with popularity, and that we can predict the location for a tweet with better accuracy and diversity than well-known baselines. We now put this into practice and evaluate our proposed algorithm that aims to filter Twitter timelines so that geographical diversity is promoted.

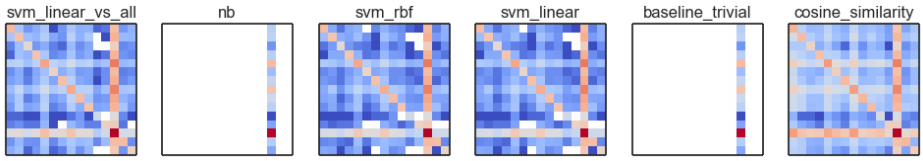


Figure 4.8: Confusion matrices from the classifiers built with location similarity features. Color encoding goes from blue (lower values) to red (higher values), and uses log scaling to showcase differences. White cells do not contain predictions.

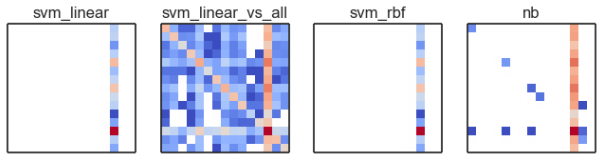


Figure 4.9: Confusion matrices from the classifiers built with *bag of words* features. Color coding is the same of Figure 4.8.

4.5 EVALUATION OF THE FILTERING ALGORITHM

As observed, timelines in Chile are less geographically diverse than expected given the population distribution. Our algorithm generates geographically diverse timelines, which, in theory, will allow users to be exposed to non centralized timelines without losing interesting and informative content. In this section we evaluate experimentally if the theoretical aspects of the algorithm hold in a centralized context according to the perception of users.

4.5.1 *Conditions and Datasets*

In addition to our *Proposed Method* (*PM*), we consider the following baseline conditions:

1. *Popularity Sampling* (*POP*): we select the s most popular tweets in terms of retweets;

2. *Diversity Filtering* (*DIV*): an implementation of the algorithm by De Choudhury, Counts, and Czerwinski [DCC11], based on maximizing information entropy of each timeline.

Since the different conditions require pairwise comparisons, and taking advantage of the low variance of geographical diversity in the dataset (see Figure 4.6), we split the dataset in:

1. *morning-noon*: 140,211 tweets published by 52,403 users between 10:00AM and 2:30PM.
2. *afternoon*: 180,824 tweets published by 63,388 users between 2:30PM and 9:00PM.
3. *night*: 401,029 tweets published by 106,942 users between 9:00PM and 2:00AM (next day).

This division of time matches the local culture where lunch happens between 1:00PM and 2:30PM, and dinner around 9:00PM. For each dataset we built filtered timelines with our *Proposed Method* (*PM*) and the baselines defined above. We excluded retweets as our focus is on standalone, source tweets.

Empirical Observations

Before the user evaluation, we evaluated empirically if our proposed algorithm generates geographically diverse timelines. From each dataset we extracted the $s = 100$ most popular tweets for *POP*, and ran *DIV* and *PM* a hundred times (*POP* runs only once because the outcome is always the same for the same input). At every timeline size $i \in [5, 100]$ we estimated:

1. Geographical diversity of *POP*, *DIV* and *PM*.
2. Jaccard similarity between *DIV* and *POP*, and between *PM* and *POP*. Jaccard similarity is defined as:

$$J(A, B) = \frac{|A \cup B|}{|A \cap B|}$$

Results are shown on Figure 4.10, with geographical diversity on the left column and Jaccard similarity on the right column. It is observed that *PM* and *DIV* consistently show greater geographical diversity than POP. In addition, *POP*'s diversity varies according to the dataset: the more tweets a dataset has, the less geographically diverse are timelines generated by POP, with decreasing averages of 0.14, 0.11 and 0.03. Conversely, *DIV* (means: 0.40, 0.40 and 0.32) and

PM (means: 0.88, 0.89 and 0.88) generate timelines with consistent geographical diversity. In fact, *PM* consistently shows greater geographical diversity than *POP* and *DIV*, and even greater than the geographical diversity of the population (0.77), indicating that our sidelining step produces the desired effect of *geo-diversification*.

In terms of Jaccard similarity, it is observed that *PM* has a consistent greater Jaccard similarity (means: 0.13, 0.12 and 0.10) with `POP` than `DIV` (means: 0.02, 0.02 and 0.01). Since we modified the `DIV` algorithm [DCC11] to start with one of the most popular tweets instead of a random selection, its Jaccard similarity with POP will never be 0. However, it can be seen that it is extremely low and, as timeline size increases, it tends to 0. In contrast, because our method considers tweet popularity, it has a consistent similarity through all datasets.

Hence, in empirical terms, our algorithm has better properties than both baselines, as we increased geographical diversity with respect to both of them, and increased representation of popular content with respect to the baseline algorithm [DCC11].

4.5.2 *User Evaluation*

According to the empirical evaluation, our information filtering algorithm generates geographically diverse timelines. By definition, our algorithm focuses on popular content, which, according to our analysis, can be geographically diverse. In this section we describe the user study we performed to evaluate how users perceive timelines generated with our *Proposed Method* (*PM*) in comparison with those of baseline conditions *Diversity Filtering* (*DIV*) and *Popularity Sampling* (*POP*). In particular, we focus on three user-centered attributes: *diversity*, *interestingness* and *informativeness*. Additionally, we group users according to their geographical origin: a centralized location (*RM*) or a peripheral one (*NOT-RM*).

Participants

Participants were recruited using *snowball sampling* in Twitter using open calls to volunteer in the study, which were retweeted by participants. No compensation was offered. We recruited 125 participants. Of them, 81 were male, 41

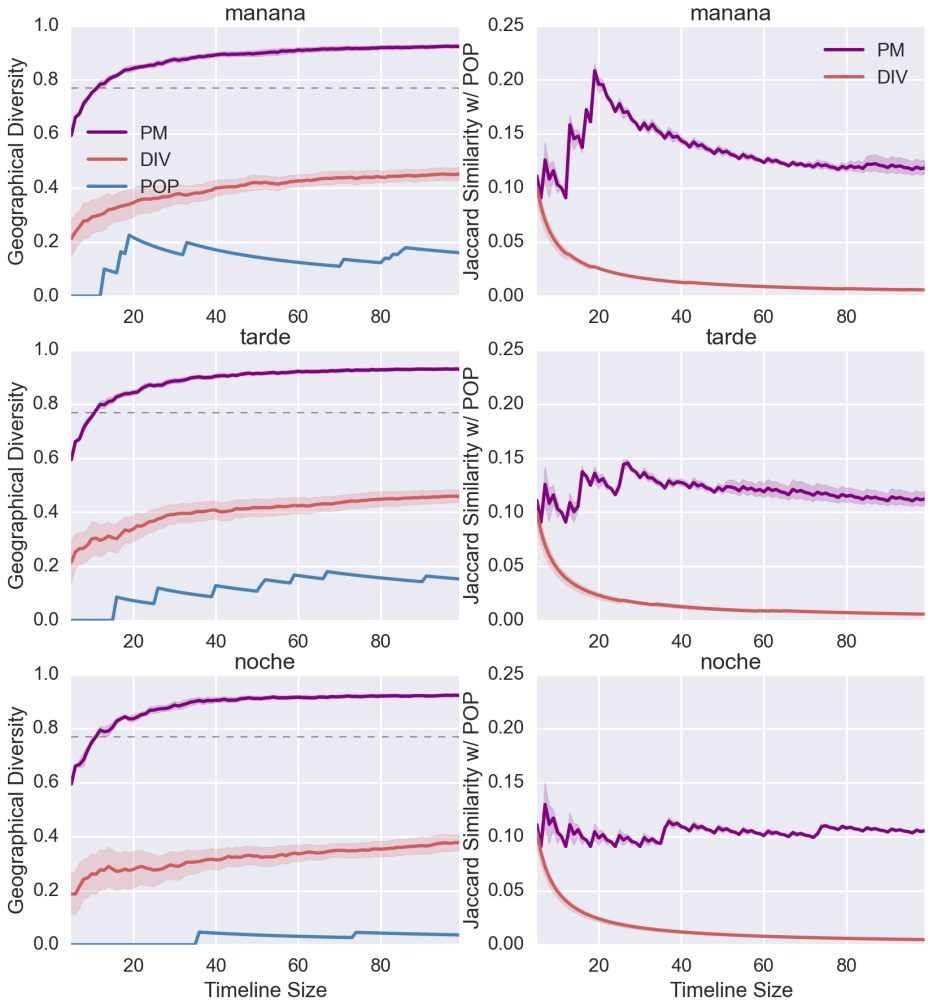


Figure 4.10: Geographical diversity for timeline sizes in [5,100] (left) and Jaccard Similarity between filtering approaches and popularity sampling for timeline sizes in [5,100] (right). All subsets of the dataset are considered: *morning-noon* (top row), *afternoon* (middle row) and *night* (bottom row). Bands represent 95% confidence intervals. Dashed lines represent the population geographical diversity (0.77).

Figure 4.11: Timelines displayed in the user study interface. Timelines rendered in this way were displayed side by side at each task from the user study.

were female and 3 opted to not say. In terms of age, 1 was 18–19, 59 were 20–29, 54 were 30–39, 4 were 40–49, 1 was 50+ years old and 6 opted not to say. All participants were from Chile (87 from *RM* and 38 from *NOT-RM*) and familiar with Twitter. Participants' experience with social networks was asked using a five-point Likert scale: those from *RM* scored 3.73 ($\sigma^2 = 0.59$), and those from *NOT-RM* scored 3.68 ($\sigma^2 = 0.58$).

Experimental Setup

For each sub-dataset, we used a multi-class SVM [CV95] classifier to predict each tweet location. Then we generated three timelines (with $s = 30$ tweets) per sub-dataset using the *POP*, *DIV* and *PM* conditions. We excluded retweets as our focus is on stand-alone, source tweets. Timelines were displayed with a format that resembled the typical Twitter user interface (see Figure 4.11). We did not display user avatars nor attached images to avoid visual distractions.

Procedure

The study had a *within-subjects* design. First, participants were asked to fill a questionnaire about demographic information and other features such as Twitter usage. Then, in at most three steps, users performed a series of comparisons between two timelines rendered side by side, each one generated by a

different condition. To avoid sequence effects, the order of pairwise comparisons (*POP/PM*, *POP/DIV*, *DIV/PM*) and the order of sub-datasets (*morning-noon*, *afternoon*, *night*) were randomized in both, the position on the screen (left or right) and the experimental step. Hence, all the participants contributed to all conditions whenever possible, as some participants were expected to not do all comparisons being an online, volunteered study. In addition, not all participants were expected to do a full read of timelines, and thus we discarded comparisons where the total reading time of both timelines was less than one minute. Note that the aforementioned situations could invalidate a counterbalanced design. Thus, we opted for a randomized instead of a counterbalanced design.

Task

Participants were instructed to read the two timelines side by side, and then answer the following questions:

1. *Which of the two timelines is more diverse?*
2. *Which of the two timelines is more interesting?*
3. *Which of the two timelines is more informative?*
4. *Optional: Please explain your answers. Add examples if needed.*

Questions 1 to 3 had a seven-point Likert scale from -3 to 3, where -3 (or 3) means that the timeline on the left (or right) is more *diverse, interesting or informative* than the other, and a value of 0 means that there was no perceived difference. Question 1 asked for general diversity as we did not want to prime participants into thinking primarily about geographical diversity, although we explained that it should be considered in its widest sense, including geography, by adding the following subtitle to the question: "*Consider diversity in its widest sense (geographical, demographical, topical, temporal, etc).*" Question 4 presented a free-text form element. After answering the four questions, a pause screen was shown for 15 seconds to allow participants to rest.

Statistical Model

The aforementioned questions define three dependent variables: *diversity*, *interestingness* and *informativeness*. For each dependent variable we built the following statistical model:

$$Y = C(\text{comparison}) + C(\text{comparison}) : C(\text{location})$$

Where C(comparison) is a dummy variable that encodes the specific pairwise comparisons performed in the study, and C(location) is a dummy variable that encodes whether users are from *RM* or not. We include an interaction term between both factors to evaluate if geographical origin influences user perception. Over this model we performed a *generalized linear models* regression with a *proportional odds model* [McC80]. This model is also known as *ordered logistic regression*, and it is used when modeling ordinal dependent variables. It extends the logistic regression model by allowing more than two categories, considering the order of the responses, and not assuming equidistant items in the Likert scale. If the statistical interaction was not found to be significant, then we performed another regression without the interaction term. Then, to validate the model, we built a null model for which we performed a likelihood ratio test.

4.5.3 *Results*

In total, participants performed 238 comparisons: 84 for conditions *POP/DIV*, 80 for conditions *POP/PM*, and 74 for conditions *DIV/PM*. Figure 4.12 showcases the distributions of pairwise comparisons between conditions by using *violin plots* [HN98] faceted according to geographical origin of participants. When reporting results we mention only those that are significant according to their p-value, considering $p \leqslant 0.05$. We identify each result as Ri, to reference it later in discussion.

Diversity

The medians of pairwise comparisons for diversity are: *POP/DIV* $= 0$; *POP/PM* $= 1$; and *DIV/PM* $= -0.5$.

The regression with interaction terms is significant (log-likelihood $= -447.21$, AIC $= 910.41$, likelihood-ratio $= 25.233$, $p = 0.0001$). The interaction term between location *RM* and condition *POP/PM* is significant ($\beta = 0.907$, 95% C.I. $[0.072, 1.753]$, $p = 0.034$). Thus, even though diversity in *PM* was perceived as more diverse than *POP*, the effect is simple due to the interaction with location: participants in *NOT-RM* scored diversity as equal between *POP* and *PM* (R1).

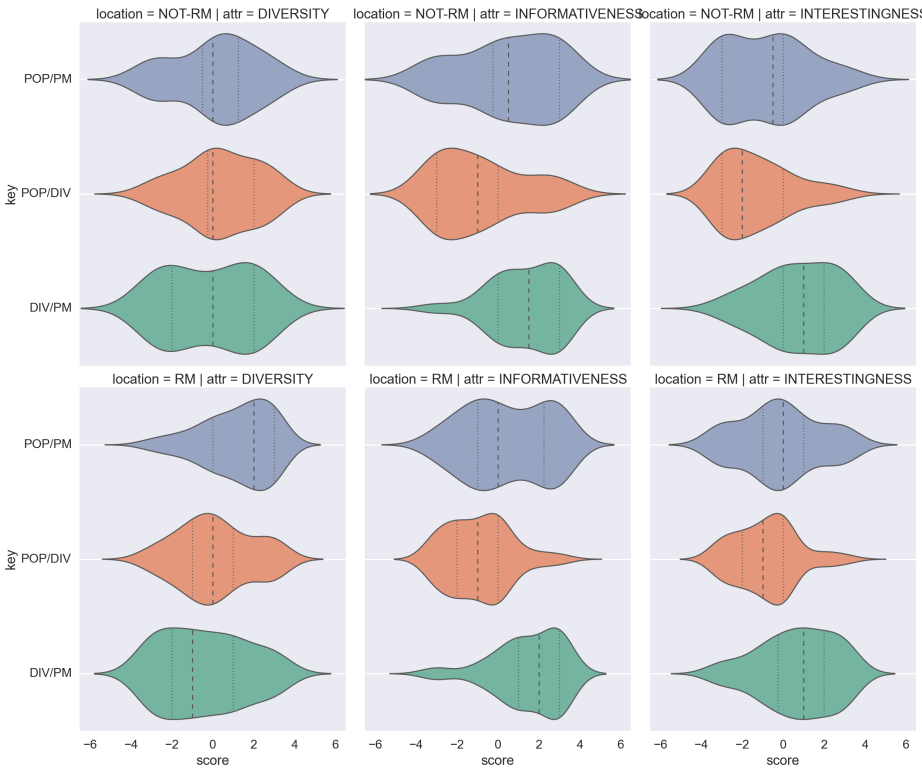


Figure 4.12: Violin plots of distributions of dependent variables diversity, interestingness and informativeness. Distributions are estimated with *Kernel Density Estimation*. A positive value indicates that the approach on the right was perceived to be more diverse, interesting, and informative than the one on the left, and viceversa. The labels of comparisons are: *DIV/PM*, *POP/DIV*, and *POP/PM*.

Informativeness

The medians of pairwise comparisons for informativeness are: *POP/DIV* $= -1$; *POP/PM* $= 0$; and *DIV/PM* $= 2$.

The regression with interaction terms is significant (log-likelihood $= -429.35$, AIC $= 874.70$, likelihood-ratio $= 49.893$, $p = 0.000$). The following terms from the statistical model are significant:

- Comparison *POP/DIV* ($\beta = -0.926$, 95% C.I. $[-1.618, -0.241]$, $p = 0.008$).
- Interaction between location *RM* and condition *DIV/PM* ($\beta = 1.364$, 95% C.I. $[0.867, 1.872]$, $p = 0.000$).

Hence, *POP* was perceived as more informative than *DIV* (R2), and *PM* was perceived more informative than *DIV* only by people from *RM* (R3).

Interestingness

The medians of pairwise comparisons for interestingness are: *POP/DIV* $= -1$; *POP/PM* $= 0$; and *DIV/PM* $= 1$.

The regression with interaction terms is significant (log-likelihood $= -430.79$, AIC $= 877.58$, likelihood-ratio $= 37.913$, $p = 0.000$). The following terms from the statistical model are significant:

- Comparison *POP/DIV* ($\beta = -1.333$, 95% C.I. $[-2.013, -0.664]$, $p = 0.0001$).
- Comparison *POP/PM* ($\beta = -0.861$, 95% C.I. $[-1.626, -0.112]$, $p = 0.025$).
- Interaction between location *RM* and condition *DIV/PM* ($\beta = 0.724$, 95% C.I. $[0.237, 1.217]$, $p = 0.004$).

Hence, *POP* was perceived as more interesting than *DIV* (R4) and *PM* (R5), although the difference is greater when comparing *DIV*. Moreover, *PM* was perceived as more interesting than *DIV* only by people from *RM* (R6).

Qualitative Feedback

The free-text answers of question 4 add a deeper understanding of how users perceived the differences between conditions. When reporting user answers, we replaced the condition names where applicable; *i. e.*, in the user comments, we would have for example *"the left column"*, which we replace by the corresponding condition, for example `DIV`. The participant numbers and their locations

are indicated within brackets (*e. g.*, [Pi] means participant i). Note that answers were translated from Spanish.

Users' answers allow us to reconstruct the motivations and design decisions made when building our algorithm:

> "I think both timelines [PM,POP] have an equilibrated set of tweets. However, in [POP] some tweets concentrate too much on only one topic: *Labbé*, *Zalaquett* and *Sabat*[2]" [P11, RM].

> "[PM] tends to show more information about on-going results in many locations of the country. The timeline [DIV] tends to show more personal opinions in the context of elections" [P36, RM].

> "I feel [PM] has much more information related to many locations, unlike timeline [DIV] which is focused on informing only about [RM]" [P50, RM].

> "[POP] is partially centralized in [RM] and the dispute between *Errázuriz* and *Labbé*[3], but it contains some analytical tweets. [DIV] is more diverse in geographical and topical terms, but [POP] is more descriptive. I liked [POP] more." [P57, NOT-RM]

> "In terms of diversity, [PM] is partly more diverse because it is more geographically diverse. I was in doubt to choose [POP] as more diverse because it contained some tweets about several topics, but at the end I preferred geography" [P76, NOT-RM].

Even though we asked for general diversity, some users explicitly mentioned centralization [P57] and geography [P36,P50,P76]. Since *POP* has very low diversity (recall the empirical observations from Figure 4.10), it is expected that geographical content to be more salient content-wise, for instance, through the appearance of local hashtags and candidate names, as noticed by Participant 57.

2 Politician names with conflictive on-going results in the election.

3 This dispute was the most unexpected one. Figure 4.1 shows both last names as some of the most popular terms.

One reason popularity was considered in *PM* was to avoid potential noise present in *DIV*, because noisy information is likely to increase entropy. This difference, as well as the similarities between both approaches, was noted by some users:

> "The biggest difference I found is how information is written and its kind. Timeline [PM] seems to be more 'formal' than [DIV](…) Timeline [DIV] shows more personal opinions" [P11, RM].
>
> "I think timeline [DIV] contained diverse opinions, not only about politics, while timeline [PM] was only about opinions on voting and election results" [P18, RM].
>
> "Timeline [DIV] is mostly opinions, in timeline [PM] opinions were better structured with concrete data" [P22, RM].
>
> "According to my taste, [DIV] contained too much trivial stuff" [P36, RM].
>
> "It seems to me that both timelines [DIV,PM] are similar in the background, although [PM] has an informative emphasis, more than merely anecdotal as [DIV]. In a way, it could be said that [DIV] is slightly more diverse in comparison to [PM], but in my judgement those 'diverse' tweets were not interesting enough, instead they are more likely to be 'noise' in the main topic of both timelines" [P47, RM].
>
> "Timeline [DIV] is more diverse [than PM] because it contains more personal opinions about the electoral process, for instance, by showing people who are not interested in voting and others who find that voting is necessary. Timeline [PM] contained factual information and trends…Therefore, [DIV] is a bit more interesting even though [PM] is absolutely more informative (data instead of opinions)" [P49, RM].
>
> "I find [POP] more interesting as it talks about a particular topic, while [DIV] is about self-referential stuff that is not always of common interest" [P86, NOT-RM].

Users agree that a noticeable difference between *PM* and *DIV* relies on the fact that *PM* presents tweets with *concrete data* [P22], *factual information* [P49], and it is *more formal* [P11]. In our context, most of popular Twitter accounts are journalists, and thus their opinions are perceived as more objective and supported than personal opinions from regular people [P86]. Participants highlighted this difference in user types in several ways:

> "Tweets from timeline [PM] were made by a bit more educated people, who had something to inform instead of publishing only personal opinions as in timeline [DIV]" [P32, RM].
>
> "In general, timeline [PM] is based on known sources or journalists. Timeline [DIV] is about people from the community giving their impressions on what is going on, which is interesting but it doesn't have any backup information" [P38, RM].
>
> "It could be said that timeline [DIV] has a diversity of users that are not, in contrast with those of timeline [PM], notable personalities of the 'Twitter world', however this is not enough for it to be considered more diverse (...). In terms of informativeness, timeline [PM] is ahead, since it is comprised of Twitter users dedicated to inform, it is expected for it to contain more complete tweets in terms of what was going on" [P47, RM].
>
> "[PM] presents tweets from people with many followers while [DIV] presents tweets from random people. Random tweets can be funny the first time you see them, but after that it's just noise." [P78, NOT-RM].

These perceptions support the quantitative results where *PM* was more interesting and informative than *DIV* by people from *RM*.

When comparing *POP* and *PM*, there does not seem to be agreement on whether which one is more interesting or informative:

> "Timeline [POP] is centered around the topic, specially in on-going results, where it contributes more than [PM]. Even though [PM] touches other topics it has more personal opinions, therefore, it is less interesting" [P9, RM].

> "Even though both [POP and PM] shared tweets, [PM] contained concrete result data, making it look more informative and objective" [P11, RM].
>
> "I have the impression that [PM] was more informal. [POP] seemed to be dominated by serious tweets." [P46, NOT-RM].
>
> "[PM and POP] are equally interesting, as they have many different perspectives about the election day. However, [POP] is more informative, because it includes different types of comments on the election day (not just news or vote counting)." [P76, not from RM]
>
> "[PM] was more informative because it contained more factual data; on the contrary, [POP] had more personal opinions" [P80, NOT-RM].

This is intriguing given that, quantitatively, *POP* and *PM* are equally informative, but *POP* was perceived as more interesting than *PM*.

In summary, qualitative feedback supports our design decisions, and partially explains the quantitative results obtained, with exception to informativeness and the statistical interactions found. Participant feedback concentrated on the users present on the timelines, as well as their content in terms of objectivity/-subjectivity, informativeness and interestingness. This focus is understandable–centralization is a systemic problem that most people is aware about, but where, at the same time, people feel they have no inference nor influence. Such behavior can be seen as a form of *conformity* [CG04].

4.5.4 *Implications of Results*

We identify two specific implications from the quantitative results.

Popularity is more valued than Diversity

In previous work by Chen *et al.* [Che+10], it was shown that popularity of content is a good feature to consider when recommending tweets. Our results are coherent with theirs, as we have found that users give more value to popularity than diversity, if we consider value as the mixture of informativeness and

interestingness. As observed with the empirical observations, *POP* has almost non-existent geographical diversity, yet it is perceived as more informative than *DIV*, which is based only on diversity (R2). This is a good result, as it indicates that *PM*, which uses a mixed approach, is equally informative as *POP*, however, *POP* is perceived as more interesting than *DIV* (R4) and *PM* (R6).

Effectiveness of our Proposed Method depends on Geographical Origin of Users

By design, we expected *PM* to be more diverse than *POP*, and more informative/interesting than *DIV*. Qualitative feedback supported our design decisions, and results R1, R3 and R6 confirmed this expected behavior by users, but *only when users come from centralized locations*, as found by the statistical interactions.

In terms of diversity, we hypothesize that people from the over-represented group (*RM*) find our approach more diverse because they are not used to see information from the outside. It is known that the geographical span of ego-networks in Twitter is small [QCC12], and thus, exposing those users to interesting and informative views from other locations expands their vision. In contrast, people from the under-represented group (*NOT-RM*) do not see differences in diversity because they are used to be exposed to views from somewhere else. Before filtering, timeline content focused prominently the centralized location; after, it contained a wider set of locations, but still not prominently their own: *"they are alike, we are diverse"* [QJ80]. Because by design our method is geographically diverse, in the next section we seek to enhance *diversity awareness* of users.

4.6 AURORA TWITTERA: DIVERSITY-AWARE DESIGN

In this section we propose an application design that seeks to increase awareness of the diversity present in timelines generated by our information filtering algorithm. This application is deployed "in the wild" [Cra+13], *i. e.*, we target end-users in their everyday use of Twitter. Because our context is not task-based, as a measure of involvement with the application we evaluate diversity-awareness through interaction events with the application and user engagement metrics [LOY14].

4.6.1 *Design Rationale*

Since timelines generated by our algorithm are diverse by definition, perhaps what people from under-represented groups need is to *become aware* of their equal representation with others in the information stream. We propose to increase this awareness by facilitating identification: *"Identification always relies upon a difference that it seeks to overcome, and that its aim is accomplished only by reintroducing the difference it claims to have vanquished. The one with whom I identify is not me, and that 'not being me' is the condition of the identification. Otherwise [...] identification collapses into identity, which spells the death of identification itself"* [But06]. Hence, we explore how to make users aware of their identity with respect to geography by making diversity visually salient.

Salience through Clustering Content and Space-Filling Visualization

Prior work determined that presentation [MR10; Par+09], nudging [MLR13] and visualization [Far+10] improve the way users behave in the presence of diverse information. In particular, the grouping of news headlines according to agreement with political positions improved user access to diverse information, measured in clicks on those headlines [Par+09]. Clustering content in locations would make easier for users to quickly see their own locations represented. However, it must be applied with care because our context is different. In political contexts, information is usually classified into a bipartite separation of groups [AG05; Con+11b], whereas in our case the number of locations is larger (for instance, 15 Chilean regions), creating the need to scroll on the screen and thus inducing a positional bias of clusters, by giving more importance to those clusters already visible without scrolling.

To avoid scrolling and its associated positional bias, we consider a previous visualization of news headlines by Weskamp [Wes04], which uses a 2D space-filling layout algorithm to partition the available screen space: the *treemap layout* [JS91; BHV00]. We visually encode locations as *internal nodes* and tweets as *leaves*. Sibling leaves appear together. The *area size* of each leaf depends on the number of retweets of the corresponding tweet, and the number of followers and friends of its author, in inverse proportions to population location. In this way, screen space is shared in an equal way between locations, as shown

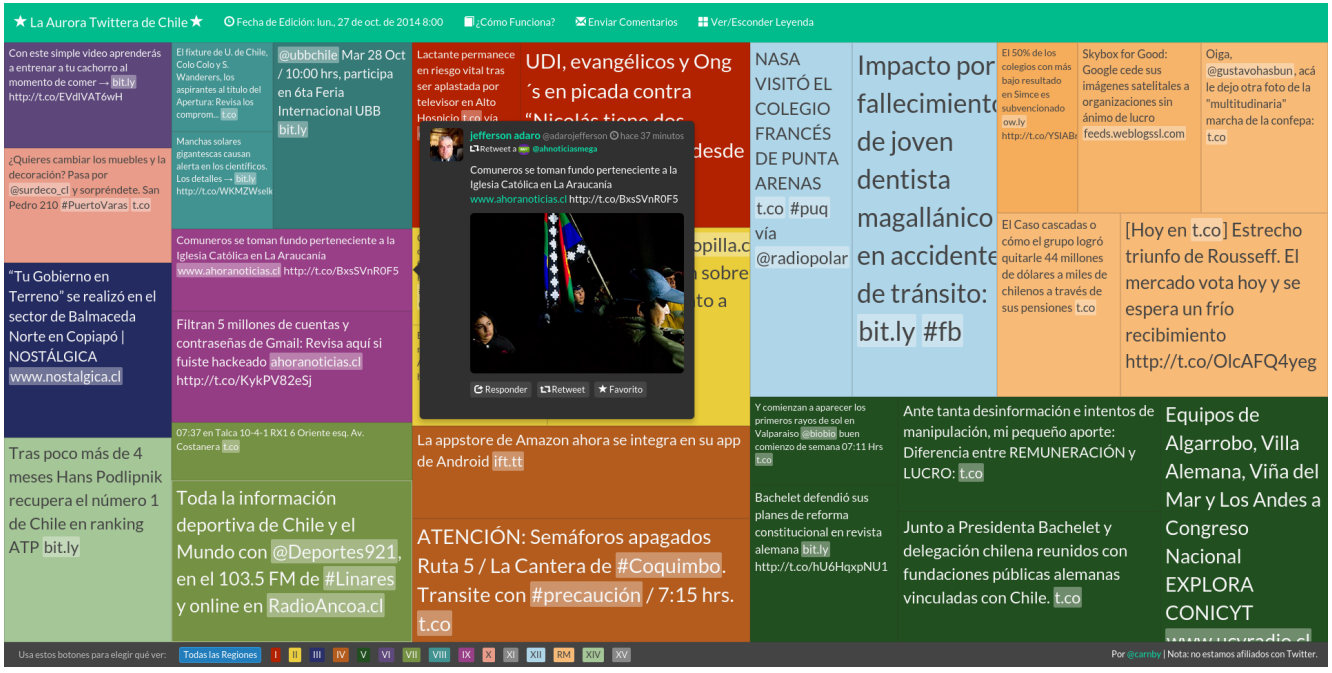


Figure 4.13: Screenshot from http://auroratwittera.cl, the URL of our prototype implementation. The bottom bar contains the location filters. The upper bar contain navigational links to an *About* page and a feedback form.

in Figure 4.13. Each leaf node is colored according to its location (*hue*) and its recency (*saturation*). Internal nodes are not displayed (they are not needed).

Interaction with Content

To interact with the visualization, users can *click* over a leaf node to display a pop-up with detailed information about the corresponding tweet, with buttons to perform core Twitter interactions (reply, retweet, mark as favorite, follow) and a text format that resembles the typical tweet presentation. To *filter locations*, for each location we display a button that, when clicked, updates the treemap to display only tweets originated from the selected location.

4.6.2 *Prototype*

We implemented a prototype of the user interface using the *d3.js* [BOH11] library. This interface, available at http://auroratwittera.cl, is displayed in Figure 4.13. It is named *"Aurora Twittera de Chile"* (*AT* from now on) as a homage to the first Chilean newspaper, *"Aurora de Chile"*. Every 30 minutes a *"new issue of AT"* was generated by the filtering algorithm (having $s = 30$ as size for each timeline and $n = 5$ for turns in the sideline step). We use the same implementation of the filtering algorithm from the first user study, with two differences: first, for performance reasons, we did not use a location classifier. Instead, we considered only tweets from accounts with a known self-reported location, although we did consider non-geolocated tweets retweeted by those accounts. Second, we avoided repeated authors or tweet content in the same timeline, and we discarded tweets where almost all text was in uppercase letters to avoid *shouting*. Input tweets were downloaded with a crawler using the *Twitter Streaming API*. As query keywords we used location names, political terms, and other terms of interest that appear constantly on the news, as well as mentions to media accounts, both at national and local levels.

Each issue had a specific URL in the form http://auroratwittera.cl/timeline/ID, which allowed users to access it at a later time, as well as saving a permanent link. In addition, if the URL contained the code of a location (*e. g.*, http://auroratwittera.cl/#RM) the interface displayed immediately the tweets related to that location in the same way as if a location filter button had been pressed.

4.6.3 *Social Bot @todocl*

Since social bots can generate social discussion and behavioral changes based on their activity [Aie+12], we created a social bot on Twitter, with username *@todocl*, to publicize *AT* and recruit users. *@todocl* presented itself as a social experiment to establish an informative community about current happenings in Chile using Twitter, and published three types of tweets:

- Whenever a timeline was generated, *@todocl* published two tweets with a link to its corresponding *issue*: one mentioning four users who authored "featured tweets", and one mentioning four users with "featured retweets" (in both cases users were selected randomly from the pool of tweets).
- After publishing those tweets, every minute *@todocl* retweeted one tweet featured in the current issue.
- Every hour past 45 minutes, *@todocl* published 15 tweets, one per location, featuring a link to the current issue with each specific location in the URL, as well as an attached image with a *wordcloud* of their most representative terms obtained with TF-IDF [BR11a].

To avoid *spamming* user timelines, at most three tweets were published per minute. Those three types of tweets can be seen on Figure 4.14.

This implementation, comprised of filtering algorithm, user interface and social bot, creates a platform where users can access geographically diverse information, in the form of an external application to Twitter, as well as injected into the platform itself. We evaluate this application next.

4.7 ENGAGEMENT AND INTERACTIONS

We study user perception in the presence of geographical diversity based on the interaction data obtained from AT. Note that AT is not a task-based system, *i. e.*, we do not expect users to visit the site to perform a specific task. Instead, the system is designed as an exploratory interface to geographically diverse timelines. As such, common evaluation metrics like accuracy and performance cannot be applied.

Figure 4.14: Example tweets posted by the social bot `@todocl`. Top: featured retweets, with a link to the current issue of Aurora Twittera. Middle: featured tweets, with a link to the current issue. Bottom: current discriminative keywords for a specific location, with a link to a specific location view in the site.

We propose that perception can be analyzed by considering metrics of interaction with the site, as well as user engagement: the number of times user interact with location filters, the tendency to return to the site, and dwell time [LOY14]. Following from the results from the first user study, we analyze if there are differences in behavior according to geographical origin, and we analyze such differences to elaborate plausible explanations of potential changes in perception.

4.7.1 *Experimental Setup*

We gathered interaction data from October 6th, 2014 until January 20th, 2015. The server logged each user request and was able to identify sessions based on cookies placed on user browsers. In total we obtained 187,604 events of the following types: *session created/restored*, *timeline and UI loaded*, *clicks* on every element on the user interface, and *pings* sent by the user interface to the server. Those events were sent through Javascript, and thus were not always reliable (*e. g.*, advanced users deactivate cookies and Javascript, and bots/web crawlers do not usually support Javascript).

User Validation

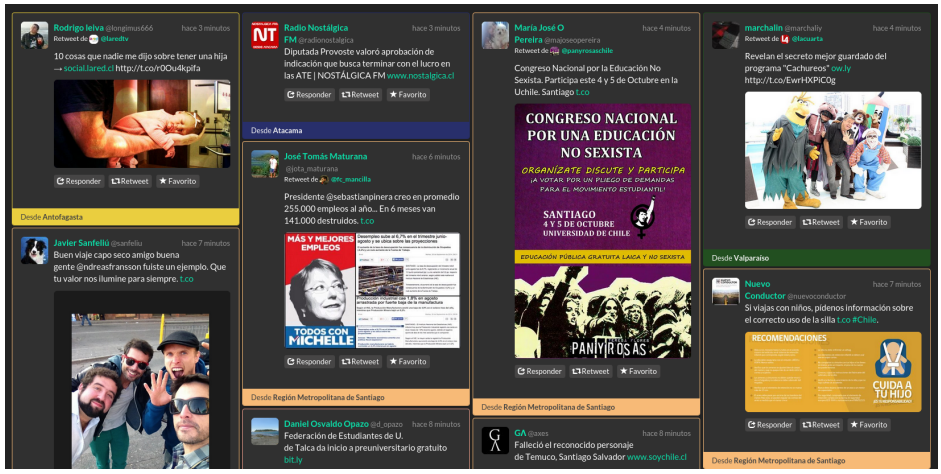
IP addresses were used to identify each user's location, using the GeoIP Legacy Database.[4] The User Agent information was used to determine if the user was using a mobile device; those users were served with a minimal version of the site but were not considered in the study because of platform heterogeneity (in terms of interaction capabilities, screen sizes, etc.). The following users were discarded: 1,660 without reliable interaction data (most of them are crawlers), 1335 with mobile devices (because our implementation is aimed at desktop users), 167 users who spent less than 10 seconds or more than 15 minutes on the site (*e. g.*, they left the browser window open), and 174 users who could not be geolocated to *RM* or *NOT-RM*. From 3,243 valid interaction events, we have 298 users, of which 173 are in *RM* and 125 are in *NOT-RM*.

Conditions

To evaluate the effect of both location and user interface on user behavior, we developed two alternative baseline conditions to compare with our own design:

1. *Baseline*, where each tweet is rendered independently of the others, and tweets are sorted by their time of publication. Each tweet is displayed inside a box with a bordered color and a legend at the bottom to indicate its originating location.

4 http://dev.maxmind.com/geoip/legacy/geolite/



(a) Baseline I: Standalone Tweets.

(b) Baseline II: Clustered Tweets.

Figure 4.15: Design baselines implemented for the study.

2. *Clustered* [Par+09], where tweets are clustered by location. Each location is represented as a box with a bordered color and a legend at the top to indicate the originating location of its tweets.

Both baselines are shown in Figure 4.15 and have the same interactive location filters as our condition, *treemap*. When a user accessed AT, if it was his/her

first visit, a random condition was assigned. Because we tracked users using cookies, in following requests users received the same interface according to their initial assignment. The distribution of users is as follows: *baseline* was assigned to 86 users; *clustered* to 89 users; and *treemap* to 123 users.[5]

Setup

We consider a *between-subjects* design, as participants were exposed to one condition only. When users loaded AT we gathered the following user information: *IP address*, *HTTP Referrer* and *User Agent*. Then, we logged each interaction with elements of the user interface. If the user requested the page through an URL with a location code, we considered it as an initial click on a location filter. Finally, the user interface sent a ping every ten seconds to the server to track the time spent on the website even in the absence of interactions, to capture the dwell time of passive users.

Statistical Model

For analysis, we focus on the following dependent variables calculated from the interaction data: *tendency to return to the site* (estimated from the *session count* by each user), *time spent* (number of minutes the user spent reading or interacting with the site, measured after the entire user interface and timeline was loaded), and *selected locations* (number of times each user filtered specific locations in the user interface). We consider the two categorical independent variables *location* (*RM* or *NOT-RM*) and *condition* (*baseline*, *clustered*, or *treemap*). Both are included in the following statistical model:

$$Y = C(\text{location}) + C(\text{condition}) + C(\text{location}) : C(\text{condition})$$

Over this model we perform *generalized linear models* regressions with the following link functions for each dependent variable: *logistic* (*logit*) for *tendency to return to the site*, *Gamma* distribution with inverse-power link function for *time*

5 Conditions are not balanced due to randomization and user validation. The ideal scenario would have been a counterbalanced design, however, because we cannot fully validate users until the end of the study, such design would have not proven to be counterbalanced in practice.

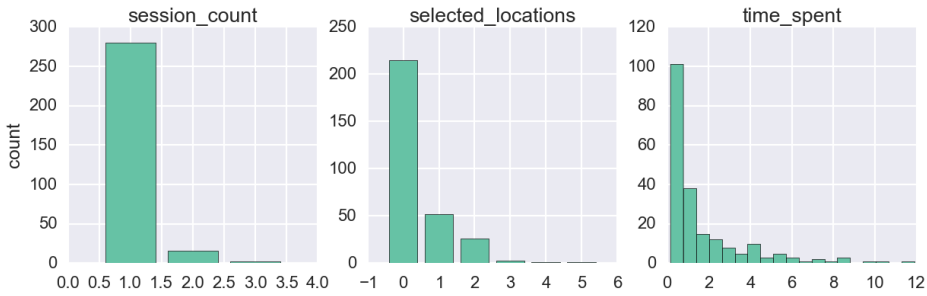


Figure 4.16: Distribution of each variable analyzed from interaction data.

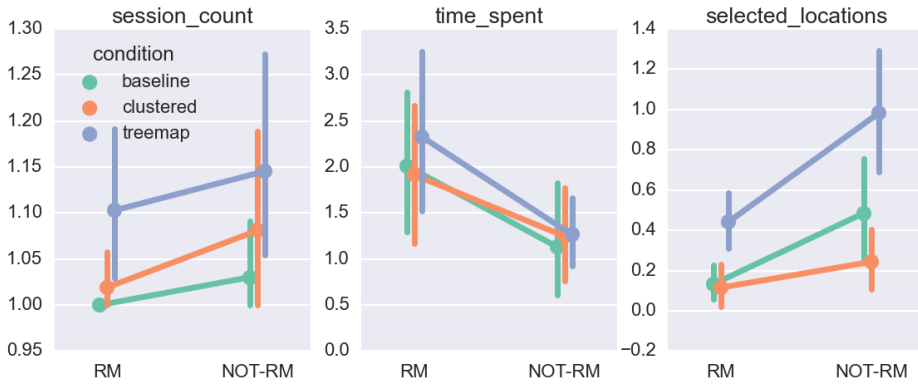


Figure 4.17: Point-plots of pairwise comparisons of interaction data for each variable between *RM* and *NOT-RM* user groups.

spent, and *Negative Binomial* for *selected locations*. For each variable, if the statistical interaction was not found to be significant, then we performed another regression without the interaction term.

4.7.2 *Results*

Figure 4.16 shows the distributions of our variables, and Figure 4.17 shows point-plots of pairwise comparisons for each dependent variable and condition between the *RM* and *NOT-RM* user groups.

Session Count and Tendency to Return

The mean number of sessions on the site was 1.07, with a maximum number of 3 sessions. No significant interaction terms were found in the first regression. In the second regression (*logit*, Pseudo $R^2 = 0.09$,log-likelihood $= -62.057$, with intercept $\beta = -4.037$, 95% C.I. $[-6.055, -2.018]$, $p = 0.000$) there is a main effect of condition (R7): users in the *treemap* condition ($\beta = 2.271$, 95% C.I. $[0.214, 4.327]$, $p = 0.03$) were more likely to return to the site a second time.

Time Spent

The mean time spent on the site is 1.70 minutes, with a maximum of 11.88 minutes. No significant interaction terms were found in the first regression. In the second regression (*Gamma*, deviance $= 267.80$, $\chi^2 = 288$, log-likelihood $= -317.36$, with intercept $\beta = 0.852$, 95% C.I. $[0.608, 1.096]$, $p = 0.000$) there is a main effect of location (R8): users from *RM* spent more time on the site than users from *NOT-RM* ($\beta = -0.348$, 95% C.I. $[-0.567, -0.129]$, $p = 0.002$).

Selected Locations

The mean number of times users selected a location is 0.40, with a maximum number of 5 times (note that this variable was normalized according to the number of sessions of each user). No significant interaction terms were found in the first regression. In the second regression (*Negative Binomial*, deviance $= 188.31$, $\chi^2 = 240$, log-likelihood $= -222.05$, with intercept $\beta = -0.951$, 95% C.I. $[-1.423, -0.480]$, $p = 0.000$) there are main effects of location and condition:

- R9: users from *RM* ($\beta = -0.833$, 95% C.I. $[-1.248, -0.417]$, $p = 0.000$) were less likely to select locations than users from *NOT-RM*.
- R10: users in *treemap* condition ($\beta = 0.848$, 95% C.I. $[0.342, 1.354]$, $p = 0.001$) were more likely to select locations than users in other conditions.

4.7.3 *Overview of Results*

We identify two main results from this study:

Behavior Differs According to Geographical Origin

The analysis of interaction data allowed us to confirm that people from *RM* behaved differently from people from *NOT-RM*. Not only they perceived content differently, as found in the previous study, they also behaved differently when interacting with the site. In particular, users from *NOT-RM* performed more location selection clicks on the user interface than users from *RM* (R9), a signal that hints that they were looking for their locations in the timeline; and users from *RM* had greater dwell time (R8), a signal that hints that they read more content than users from *NOT-RM*. This might mean that users from *NOT-RM* looked for the content related to their locations (and nearby ones), and read that content only, while users from *RM* did not filter locations, but read much more content–they were less focused in terms of geography.

Treemap helped to Increase Access to Diverse Content

The baseline conditions did not have any significant effect on the performed regressions. However, the treemap condition was found to increase likelihood of returning to the site (R7), as well as encouraging users to seek for specific content regardless of their geographical origin (R10). Surprisingly, the text-based clustered representation did not encourage access to diverse content, as expected given the results of Park *et al.* [Par+09].

At the beginning of this section we stated that the main purpose of AT's design was to make users aware of diversity. We believe we succeeded, because, regardless of location and in addition to the positive engagement, the treemap design makes users explicitly select more locations than with the baseline user interfaces.

4.8 DISCUSSION

Our results suggest that in micro-blogging platforms the entire life cycle of content generation and consumption is affected by systemic biases. In this section we discuss the implications of the case study, user evaluation and in-the-wild results of the work presented in this chapter.

4.8.1 *Implications*

Biases from the Physical World affect Content Processing

It is expected that virtual platforms are affected by physical constraints [TGW12], however, as predicted by Gillespie and Robins [GR89], our case study shown that, even when lacking geographical barriers, the population behaved in a centralized way, by making the information flow biased towards the central location in comparison to a non-biased flow based on population distribution. In turn, this biased behavior influences content-based algorithms, as shown in our study of machine learning classifiers that tend to over-fit when diversity is not considered, although they do not lose accuracy. Hence, even when in theory a content-based algorithm is bias free, its input content might not be, and this is hard to find if accuracy is the only metric being optimized.

Because humans can be prone to accept machine-generated content and decisions, a bias known as *automation bias* [Cum04], diversity should be considered by system designers in the entire content-processing pipeline.

Algorithms are not Enough to Encourage Exploration of Diverse Timelines

Algorithms are not enough because, if the user interface is not designed to surface the differences that shape identity [But06], then users will not be aware of the diversity provided by the algorithm. In this aspect, our study "in the wild" helped to understand if centralization shaped behavior of end-users when interacting with a system designed to encourage access to diverse information. The results obtained imply that, indeed, visualization techniques like *treemaps* [JS91] help users to *see* the inherent diversity on timelines filtered by our algorithm. In particular, we based our idea in the design by Weskamp [Wes04],

which has proven to be useful in this context as it *"allows many interesting comparisons and readings of how we differ culturally"* [Mei13].

User interfaces in Web platforms are still focusing on business-like interactivity and aesthetics [Don14]. Our results imply that interaction designers should consider other interfaces that would give people the tools needed to explore content in new, and hopefully unbiased, ways.

Social Contexts and Individual Differences

Information seekers and content explorers are affected by systemic biases, even when they are not aware of it. As we observed on the first user study, perception with respect to geographical origin (in terms of centralization) was affected by centralization, as users did not see the diversity present by definition in our generated timelines. When analyzing qualitative feedback, user answers allowed us to reconstruct the design decisions behind the information filtering algorithm. However, diversity was not perceived quantitatively, even though it was mentioned in answers. We hypothesize that this is due to conformance with centralization [CG04], although this needs to be evaluated qualitatively in further studies.

In the literature, cultural differences [HHM10] have been acknowledged in the study of communication systems [KFS06] by suggesting specific features and interaction mechanisms pertinent to each culture. Because we studied anonymous end-users, we were not able to study specific cultural differences, although we acknowledge their importance. In contrast, we identified individual differences based on whether users belonged to central or peripheral locations in the context of political centralization, and found that this distinction explained the differences in behavior found in our study "in the wild". This implies that differences based on systemic biases should be considered when designing end-user systems, either to provide personalized content (but care must be taken to avoid biased personalized content), or adaptive user interfaces with specific interaction mechanisms suitable for the social context of the user.

4.8.2 *Summary, Limitations and Future Work*

In his essay *"In praise of shadows"*, Jun'ichirō Tanizaki wonders what if the fountain pen, an *"insignificant little piece of writing equipment"*, would have been invented in Japan, as in spite of its insignificance it *"had a vast, almost boundless, influence on our culture"* [Tan01]. In that line of thinking, we wonder "*what if global Web platforms would have emerged in countries with severe systemic biases?*" Perhaps algorithms would have considered those biases and user interfaces would have been adapted to mitigate their effects. As we have found in our work, this is possible by having a *pluralist design* [Bar10] approach to study perception, engagement and behavior of users. To this end, we defined a methodology to understand if, and how, centralization is reflected from the physical world into the virtual population of a micro-blogging platform, as well as to promote geographically diverse content. We studied the specific case of Chile, a highly politically centralized country [GK08], and validated the rationale of our methodology. Then, with Chilean users, considering their geographical origin from a centralization point of view, we analyzed through carefully designed experiments the differences in users' perception of diversity when exploring geographically diverse timelines.

In a user study with labeling tasks, we found that, while users from different locations in Chile agree on which content is informative and interesting, only users from the centralized location were able to see the diversity present in the constructed timelines. Then, inspired by a visual design by Weskamp [Wes04], we used information visualization to make users *aware* of such diversity, addressing what we called the *diversity-awareness* problem. We deployed this design on the Web, and spread information about it using the social bot *@todocl*. By analyzing logged interaction data, we observed that users behave differently according to their geographical origin, but also that an ensemble of information filtering and user interface design can improve exposure to diverse information, at least in the short term, as our study does not allow us to draw conclusions on the long term.

When analyzing behavior quantitatively, we did not focus on effect sizes of statistical differences. We rather focused on the presence or absence of differences in behavior, explained through statistical models, in particular *generalized linear models*. Considering that, as of April 2015, Twitter has 288 million

monthly active users, with 77% of accounts outside of the United States[6], we believe that consideration of our implications, even on the presence of small effect sizes, would have noticeable, and hopefully positive, consequences on information access by users.

As mentioned in the introduction, social contexts must be considered when designing systems [Bux10]. Our work has shown that, in addition to culture, systemic biases also shape social behavior and perception, and thus, they must be accounted for in system (algorithms and user interface) design and evaluation.

Limitations

Critics might rightly say that we did not control user sampling on our experiments based on other sociodemographic characteristics than location. Currently, given centralization and the unequal population distribution, finding users from *RM* is easier, because its population is greater and has more access to Internet in comparison to other locations, making it harder to find users willing to participate from non-central locations. In this aspect, our snowball sampling method provided a way to find a large enough population to gain important insights in the first study. In the second study, this limitation does not hold, as all users mentioned by *@todocl* were geographically diverse, and those users retweeted *@todocl*'s tweets, improving the representativity of the sample.

Moreover, while we have performed a quantitative evaluation of user behavior with our system design, there is still a need to understand the *why* behind the differences in user engagement and interaction behavior. This can be explained only with qualitative studies.

Future Work

The Web is increasingly being more accessed from mobile devices than from desktops. In our case, of the detected mobile users, 68% could be geolocated using the IP address. Considering those users in a new mobile-friendly version of our site will allow us to study differences in a mobile context. Ideally, this would be a longitudinal study with a qualitative component, which will also help us to

6 https://about.twitter.com/company

address the limitations of this paper. We will also study response to the social bot *@todocl*. We created it as a way to drive traffic to our site, with satisfactory results, but we also noticed that some users started to reply and retweet our retweets, a behavior that could be analyzed in terms of relevance feedback. Finally, we look forward to expand our work to other diversity scenarios, to see if our design process and evaluation allow users to be exposed to more diverse information.

5

ENCOURAGING EXPLORATION WITH DATA PORTRAITS

In micro-blogging platforms, people can connect with others and have conversations on a wide variety of topics. However, because of homophily and selective exposure, users tend to connect with like-minded people and read agreeable information only. Motivated by this scenario, we propose a recommendation algorithm to suggest new people to connect with. This algorithm is focused on recommending politically diverse people, yet at the same time it makes use of homophily to find those users. We introduce a paradigm to present these recommendations injected in a *data portrait* of users, in which their user interests are visualized. To evaluate our proposed approach, we first conducted a case study on Twitter, considering the debate about a sensitive issue in Chile, where we confirmed homophilic behavior in terms of political discussion. Then we evaluated a first algorithm and design proposal, where we found that politically-vocal users had different perceptions of recommendations. Using qualitative feedback from the pilot study, we improved both algorithm and data portrait design. Their new implementations were integrated into the *Aurora Twittera* platform to find end-users who created their own data portraits. Finally, we analyzed interaction data from the usage of the platform. Our main results are: 1) *informational* and *behavioral individual differences* influenced user interaction with the system; 2) visualization increased exposure to recommendations regardless of the recommendation algorithm, meaning that visualization encourages exploration of recommendations of politically diverse people inside data portraits.

5.1 INTRODUCTION

Users have less barriers to communicate in the Web today, as geographical distance is no longer a limitation to interact with others or to know what is happening anywhere and anytime. Real-time streams in social networks and micro-blogging platforms keep people aware of what is happening simply by reading content posted from followed accounts into their timelines, as well as content from non-followed accounts that use specific hashtags of relevance for current events. Furthermore, some micro-blogging platforms have been claimed to be socio-political tools that have aided revolutions like the Arab Spring [Lim12].

However, is it really that good? Social research has shown that, while everyone indeed has a voice, people tend to listen and connect only to those of similar beliefs in political and ideological issues, a cognitive bias known as homophily [MSC01]. This kind of behavior happens in many situations, and it can be beneficial, as communication with culturally alike people is easier to handle. However, the consequences of homophily in ideological issues are prominent, both off and on-line. On one hand, groups of like-minded users tend to disconnect from other groups, polarizing group views. On the other hand, Web platforms recommend and adapt content based on interaction and network data of users, *i. e.*, who is connected to them and what they have liked before. Because algorithms want to maximize user engagement, they recommend content that reinforces the homophily in behavior and display only agreeable information. Such reinforcement, in turn, makes computer system to recommend even more polarizing content, confining users to *filter bubbles* [Par11].

Motivated by this scenario, in this part of the dissertation we approach the following research question:

> *How to encourage exposure to diverse people from a ideological point of view in micro-blogging platforms?*

Until now the literature has focused on how to motivate users to read challenging information or how to motivate a change in behavior through recommendation systems and display of potentially challenging information. This "direct" approach has not been effective as users do not seem to value diversity or do not feel satisfied with it, a result explained by *cognitive dissonance* [Fes62], a

state of discomfort that affects persons confronted with conflicting ideas, beliefs, values or emotional reactions. Conversely, we propose to follow an indirect approach, where we take advantage of partial homophily to suggest similar people, where similarity is estimated according to intermediary topics. We define intermediary topics as non conflictive shared interests between users, *i. e.*, interests where two persons of opposing views on sensitive issues could communicate and discuss without facing challenging information in a first encounter. According to the primacy effect in impression formation [Asc46], *"first impressions matter"*, making such intermediary topics important. In this way, recommendations based on intermediary topics indirectly address the problem of being exposed to people of opposing views.

We propose that context and presentation of politically diverse recommendations is also important. The proposed context is a visual depiction of the user profile called *data portrait* [Don+10], with the purpose of making users aware of their own user interests and the image they project on the social platform. The proposed representation of recommendations is based on a hierarchical visualization technique to display how *recommendees* can be grouped. Both concepts allow users to contextualize recommendations according to their interests and self-image [Gof59] projected through their portraits.

We first validated our proposal by performing a case study on the microblogging platform Twitter, with users who discussed sensitive issues, *i. e.*, ideological or political themes that would make people reject connecting or interacting with others. Specifically, we estimated user stances with respect to the sensitive issue of abortion in Chile in the context of the on-going campaigns of the presidential elections in 2013. Abortion is a good issue to base analysis on, as it has specific, identifiable stances, namely *pro-life* and *pro-choice*. Moreover, Chile has one of the strictest abortion laws in the world [Uni; SB07], yet at the same time a majority of population is in favor of its legalization [CEP13], making it a controversial topic.

The case study confirmed the homophilic structure of discussion, and it surfaced the characteristics of content and users who participated on the debate as well. Then, we built a prototype data portrait to evaluate, in a pilot study, how users perceive recommendations injected in data portraits, and found that users who have tweeted about abortion before the study had a different perception of recommendations than users who had not, in addition to deep qualitative feed-

back about the system as a whole. Following the results from the pilot study, we developed a formal definition of *intermediary topics*, *i. e.*, those topics about non-confronting, shared interests, as well as a refined recommendation algorithm and a new data portrait design.

Finally, we incorporated the refined data portrait application into the *Aurora Twittera* platform introduced in the previous chapter. Likewise, we made use of the social bot *@todocl* to help to disseminate the application. We analyzed interaction data of end-users with their own portraits and the injected recommendations. Because our application lies in the field of *Casual Information Visualization* [PSM07], it does not consider specific tasks to be performed by users. Such systems are hard to evaluate, and thus, the interaction data is analyzed in the context of *user engagement* [LOY14]. Our results contribute interesting insights:

- Usage of visualization to depict recommendations alongside a data portrait encourages users to explore more recommendations, regardless of the algorithm used to generate them.
- Recommendation acceptance is not influenced by visualization. Instead, it is influenced by political involvement. In this aspect, users behave in homophilic ways.
- *Informational behavior* [NBL10] in conjunction with political involvement influence how users engage with the system.
- From the analysis of user engagement metrics, in particular *dwell time*, we identified two types of exploratory behavior: *focused* and *reflective*. Visualization and intermediary topics encouraged politically involved users to perform a conscious decision-making process in terms of recommendation exploration and acceptance.

Hence, the effectiveness of a system like ours will depend on whether the user is expected to explore recommendations or not. This has implications on designing visual user interfaces to explore user generated content, as a one-size-fits-all approach misses the opportunity of giving users tools to get the best out of their exploring experience. We discuss the implications in terms of *who* can be targeted with systems like ours, *when* to consider presenting diverse recommendations, and *how* to engage users with data portraits. The *why*, which needs to be studied qualitatively, is left for future work.

5.2 BACKGROUND

The work presented in this part of the dissertation spans several research areas. We discuss these in relation to our aims, the positioning of our work, and approaches adopted.

Homophily and Content Recommendation

Homophily is the tendency to form ties with similar others, where similarity is bound to many factors, from sociodemographic to behavioral and intrapersonal ones (see a literature review by McPherson, Smith-Lovin, and Cook [MSC01]). In Web platforms homophily is present in sharing behavior in social curation platforms [Cha+14], interactions with others with similar emotions and linguistic style in on-line collaboration [Ios+14], and links between political parties in blog networks [AG05], among others. In micro-blogging platforms, the presence of homophily in how individuals interact allows to use their ego-network structure to predict user attributes [ALR12; Rou+13] as well as to recommend people to interact with [Che+09; HBS10]. In our particular context, it has been observed in terms of political leaning in micro-blogging platforms [Bar15].

In this chapter, we focus on recommendations in micro-blogging platforms, which, as mentioned above, are influenced by user similarity. Thus, care must be taken when defining similarity. When considering attributes similarity is clearly identifiable: people are or are not from the same location or from the same ethnicity, and they went to the same school or they did not. However, interest-based similarity is not completely defined, as there are many ways to define if two users are similar or not. For instance, two users might share interests if they use the same *tags*[1] [BR11b], follow the same accounts [Goe+13], mention the same entities [MM10], or if they have similar *latent topics* [RDL10; QAC12] estimated with Latent Dirichlet Allocation [BNJ03][2].

When recommending information and others to follow, similarity is not the only important signal. Other relevant signals include content quality and popularity [Che+12], network relevance (*friend of a friend*) [Che+09], explainabil-

1 Also known as *hashtags* in micro-blogging platforms.

2 Note that, while *tweets* are too short to be reliable for topic modeling, the concatenation of tweets into a user document is good enough.

ity [HKR00], and centrality measures [Gup+13]. Yet, all of them are influenced by similarity–someone may be popular, but if it is not similar enough to the user, it is unlikely to be interesting for him/her.

The *filter bubble* phenomena [Par11] is related to homophily, as it involves systems that filter information that might not be inline with users' interests, specially in challenging settings like political leaning. Since opposite or challenging information is filtered out, the user is unlikely to interact with opposing others or opposing information, introducing a snowball effect. However, if challenging information is not filtered out, users would still prefer agreeable, non-challenging information [LF13]. One explanation to this behavior is provided by the *cognitive dissonance* theory by Festinger [Fes62], which states that, when individuals are confronted with opposing information, they experience an uncomfortable state of mind that can be alleviated by discarding the information or avoiding it [Har+09].

In this setting, one way to improve recommender systems is to include diversity in their definition and evaluation. Accuracy is important, but there are others metrics like coverage, confidence, novelty and serendipity [Her+04], which are sometimes left out of evaluation [MRK06] and algorithm design [Abb+09]. One way to improve coverage and encourage serendipity is by ensuring diversity in recommendations. This can be achieved by minimizing similarity between items in a recommendation list [Zie+05], maximizing *information entropy* [Jos06] over a set of content features [DCC11], as well as applying context-specific diversification methods [MZR09]. Diversified sets have increased user satisfaction in book recommendation scenarios [Zie+05] but not in political scenarios unless users are *diversity-seekers*, usually a minority of users [MR10].

In our work, we propose *intermediary topics* as a feature to consider when recommending users to follow. The intuition behind intermediary topics is that they focus on homophily in specific shared interests which are non confronting nor challenging, *i. e.*, unlikely to provoke cognitive dissonance. Our definition of intermediary topics is based on topic modeling like Ramage, Dumais, and Liebling [RDL10]. However, instead of estimating user topics and estimating similarity directly, we build a *topic graph* of relations between latent topics, and find which ones are more likely to include people from diverse political backgrounds by estimating information centrality [BF05]. These graphs have been used in the past. For instance, Gretarsson *et al.* [Gre+12] visualized topic graphs

to ease knowledge discovery by analysts. We work in a different context and do not visualize them. Instead, we consider a subset of their nodes (the intermediary topics) as input features for a recommender system. When facing users, we experiment with visual depiction of recommendations, because visualization of social recommendations has increased user satisfaction in the past [Gre+10].

Exposure to Diverse Content

Exposure to agreeable information only, as well as like-minded people exclusively, reinforces and polarizes individual and group stances on ideological issues [ML75; Sun09]. In the literature, previous work has focused on how to minimize or avoid cognitive dissonance to improve exposure to challenging information, by employing algorithms for content selection as well as changing depiction of this kind of information (for a extensive review on this subject, see the doctoral thesis by Munson [Mun12]).

We identify four areas of research in terms of exposure to diverse content: *exploratory interfaces of diverse information*, *explicit diversification of the information space*, *visual cues in mainstream user interfaces*, and *indirect/implicit diversification*. Exploratory interfaces allow users to browse several stances and points of view of political issues, with the intention to compare pro/cons of each stance, and inform users. Faridani *et al.* [Far+10] presented *OpinionSpace*, a self-organizing interactive visualization of the information space, where individual opinions of participants in debate were visualized according to their *opinion profiles*, built automatically for each participant after answering questions about key political issues. Although the usage of a visual approach did not reduce selective exposure, it generated more engagement than baseline text-based interfaces and users were more respectful with those having opposite opinions. Kriplean *et al.* [Kri+12] presented *ConsiderIt*, a platform for public deliberation of political issues. In ConsiderIt, participants craft their own positions, and then are able to browse aggregated discussions of key political issues. Unlike OpinionSpace, it is not based on novel visuals. It focuses on identifying pros and cons of specific issues rather than on a multidimensional spectrum. In both applications, one assumption is that users are seeking for political information and are willing to discuss about it (not necessarily with people with opposing views).

Explicit diversification means that challenging content is directly displayed as such. In *NewsCube* by Park *et al.* [Par+09], several automatically determined aspects of news stories in political contexts are presented to mitigate the problem of media bias and allow users to access diverse points of view of news events in political contexts. Each aspect is displayed in its own cluster, allowing users to see the diversity of available points of view. This clustered presentation augments the number of interactions with news, but not the number of interactions with different, opposing, clusters [CR13a]. An *et al.* [An+14] propose an interactive news aggregator for micro-blogging platforms, where news are visualized according to four dimensions: *gratification*, *selective exposure*, *socialization*, and *trust & intimacy*. The user is responsible for establishing which dimensions are more important for him/her. In both scenarios, the user is visually aware of the many different choices because the user interface (which is different to those him/her is accustomed to) ensures their representation, but there is no incentive to have a diverse information consumption behavior.

The usage of visual cues inside mainstream user interfaces has also been developed before. Munson and Resnick [MR10] tested different ways of altering a user interface without changing its core interaction mechanisms, by changing sorting order of information as well as highlighting items pertaining to opposite points of view with respect to the user. It was found that only a minority of users, called *diversity-aware*, values diversity. Munson, Lee, and Resnick [MLR13] developed a browser extension where users were presented with a visual representation of their reading behavior of news outlets in terms of political diversity. The representation was updated every time the user read a news article on the Web and, in a subtle way, it encoded balance in behavior. This mechanism helped to balance reading of left-winged users, but not of right-winged ones. In discussion forums, Liao and Fu [LF14] added *position indicators* of stance polarization to participants, improving agreement of users with those of opposing views when their positions were not consistently moderate, or when the information seekers were looking for highly accurate information.

The indirect approach happens when users are not aware of diversification. For instance, in news aggregators the *Sidelines* algorithm by Munson, Zhou, and Resnick [MZR09] iteratively excludes articles from specific point of views until there is balance of stances in the delivered news. However, this method requires a knowledge base with point of views, and a classifier to categorize news arti-

cles into those views. Algorithms based on information entropy of content and its author, like the one proposed by De Choudhury, Counts, and Czerwinski [DCC11], can show diverse political information, provided that political leaning is one feature used in entropy estimation. Another example of indirect approach is query augmentation by Yom-Tov, Dumais, and Guo [YDG13], where specific biased queries submitted to a search engine returned results from those queries, plus results from an unbiased or even opposite version of each query, without indicating this augmentation on the user interface nor in the search results meta-data.

In our work, we apply an indirect approach, where we use intermediary topics to recommend people with potentially opposing views. Conversely to the discussed approaches, the context of our recommendations is not related to politics nor sensitive issues; instead, we build a data portrait of users of micro-blogging platforms, and show recommendations in that context, emphasizing the similarity of recommendations with the target user.

Information Visualization

Visualizations of micro-blogging data cover a wide range of applications, like event monitoring [Dor+10; Mar+11], visual analysis [DNK10], group content analysis [Arc+11], information diffusion [Vié+13] and ego-networks [Hb05]. To visually represent user profiles we consider *data portraits* [Don+10], which are *"abstract representations of users' interaction history"* [XD99]. These portraits have been built using content from e-mail [VGD06], personal informatics systems [AD09], Twitter profiles [Dra09] and discussion forums [XD99].

Our work is related to the field known as *Casual Information Visualization*, defined by Pousman, Stasko, and Mateas [PSM07] as *"the use of computer mediated tools to depict personally meaningful information in visual ways that support everyday users in both everyday work and non-work situations"*. The focus on everyday situations imply that there does not need to be a concrete task to be completed, nor a specific analytic insight to be expected.

Yi *et al.* [Yi+08] identify four cognitive processes that leads to insight: *provide overview*, *adjust*, *detect pattern*, and *match mental model*. To provide overview of profiles, as well as to match mental models of portrayed users, we use *wordclouds* as primary element of our proposed designs. Worclouds have a long

history in information visualization, as described by Viégas and Wattenberg [VW08]. Although arguably not adequate for analytical tasks, they are familiar and popular with users, as they help them to express themselves [VWF09]. We make use of wordclouds both as an overview of a profile, as well as a navigational tool to explore it, in a coordinated view with the other visual elements of the portrait.

Other visualization techniques exist to depict structure in text, like *WordTrees* by Wattenberg and Viégas [WV08] and *PhraseNets* by Van Ham, Wattenberg, and Viégas [VWV09]. Even though we model data portraits as bipartite graphs between user interests (keywords) and user generated content (micro-posts), we do not focus on relations between words nor text structure, and thus, we use *wordclouds* instead of PhraseNets.

Visualization and graphic techniques used in recommendation contexts include: controllable Venn-diagrams [PB15], network graphs [Gre+10; Gre+12], dust and magnet [An+14], and compound graphs [Gou+11]. However, they are targeted at expert users that know how to control such visualizations, or are task-based systems. Instead, in a similar way to the *Hax* application [Sav+14], we propose to use circle packing, a hierarchical visualization technique [CS03], which allows us to create a casual user friendly depiction (*i. e.*, direct and uncluttered) of our generated recommendations without needing user-controllability nor user expertise. In contrast to Hax, instead of building a user interface to find audiences to broadcast information, our user interface is aimed at finding people to interact with.

Political Leaning in Social Media

To study political leaning in social media, in particular in micro-blogging platforms, the first challenge is to actually detect which is the political leaning of users, as this attribute is not usually part of a public profile. Manual coding of annotations by experts is probably the best method to identify the political leaning of people in social networks, but this approach is not scalable. One way to address the issue of classifying users is through supervised machine learning [Con+11a; PP11], Bayesian estimation [BKY13; Bar15], and political score propagation of known sources [GH11]. Features used in classification include vocabulary, hashtags, and connectivity with accounts with known political leaning.

However, obtaining reliable results is specifically difficult with micro-blogging platforms because, unlike classification efforts for speeches [TPL06], micro-posts are short and with arguably bad lexical quality in comparison to knowledge sources like Wikipedia and Quora [RB12]. Furthermore, note that because regular people may not be as vocal as politically involved people, the accuracy of prediction algorithms is often over-estimated [CR13b].

Knowing political alignment of users allows to study group polarization. For instance, Adamic and Glance [AG05] found that liberal and conservative blogs in the US linked mostly to blogs of the same political party. Yom-Tov *et al.* [Yom+12] studied the interaction between two opposite groups of users on Flickr[3], one related to *pro-anorexia* and the other to *pro-recovery*. In Twitter, group polarization has been observed not only in terms of general political leaning [Con+11b], but also in terms of specific controversial political issues [HMS13] and religion [WGB13].

Conover *et al.* [Con+11b] found that the *mention network* in Twitter, although polarized, it is less so than the *retweet network*. However, a mention in Twitter *per se* is not a meaningful interaction, as nothing ensures that the original tweet is read or replied. Moreover, the authors indicate that content injection through hashtags is common. Injection of political content into opposing timelines can be seen as a method to reduce polarization, but it causes the opposite effect [Con+11b; Yom+12].

In a work related to our case study, Yardi and Boyd [YB10] studied debates about abortion in Twitter, in particular between users of *pro-life* and *pro-choice* stances. Their results indicate that the interaction between users having the same stance reinforced group identity, and discussions with members of the opposite group were found to be not meaningful, partly because the interface did not help in that aspect. As noted by the authors, people hardly changed position on abortion based on discussions on Twitter. However, being connected to a more diverse group of people may help create a more meaningful discussion and help people to diversify their points of view instead of merely reinforcing them. This is what we strive for in our work.

3 A social network about photography, http://flickr.com.

5.3 INITIAL METHODOLOGY AND DESIGN

The aim of this part of the dissertation is to build a tool that recommends microposts to a target user, from authors who hold opposite views in sensitive issues. As platform for study we consider the micro-blogging platform Twitter, where each user is able to *follow* other users, making their tweets available in his/her own *timeline*. However, because of homophily [MSC01], such connections are biased because of the tendency of people to connect with like-minded individuals. Moreover, when recommended with information that challenges current beliefs, users discard or avoid that information, a behavior known as *selective exposure*.

In this section we introduce our first approach to this problem, where these recommendations are injected into a data portrait of target users. We present our initial methodology to model users, in both general topical interests and specific stances on sensitive issues. We need to determine, for each user, what are his/her views with respect to the sensitive issues under consideration, and what are him/her interests in non conflictive settings, like sports, dining, film, music, and so on. Then we define a proof-of-concept recommender system that takes user interests and stances, and generates diverse recommendations from ideological point of views, but exploiting user similarity in non challenging areas– *i. e.*, we exploit homophily from *shared interests* on non conflictive topics. To contextualize those recommendations in terms of user interests, they are injected into a data portrait of the target user. This approach to recommending people of opposing views is indirect, because even though the recommended content comes from people of opposing views, the content itself is non challenging in terms of the target user's beliefs and ideas regarding sensitive issues.

5.3.1 *Sensitive Issues and Shared Interests*

Sensitive issues are political or ideological topics for which their stances or opinions tend to divide people. This considers topics like *global warming*, *social security*, *health care reforms*, and *abortion*. Such topics tend to polarize people, *i. e.*, users who support one stance in abortion do not interact with users who support another stance, and when they do, it is not a meaningful interaction.

Conversely, *shared interests* are topics for which their stances or opinions do not, in normal conditions, tend to divide people. As example, people who support the soccer team *F.C. Barcelona* has a rivalry with people who support *Real Madrid F.C.*[4], however, the selective exposure mechanism would not be activated when discriminating information coming from people who support the opposite team–in fact, in some cases, they might be interested in such information. Other contexts can be less challenging as there might no be explicit rivalries. For instance, people with different musical tastes might be interested in discussing the particularities of their liked styles for comparison with others.

5.3.2 *Representation of User Stances in Sensitive Issues*

An assumption we make with respect to user stances is that they are linked by partisan political ideology, *e. g.*, conservative/liberal people share views on different sensitive issues. Then, to estimate user stances, we first need to be able to estimate what users say with respect to to sensitive issues. In Twitter, often users annotate their tweets with *hashtags*, which are text identifiers that start with the character **#**. For instance, *#prochoice* and *#prolife* are two hashtags related to two abortion stances, and each one of those stances has specific hashtag and words related to them (*e. g.*, *"right to choose"* is pro-choice, and *"it is life since conception"* is pro-life). Pennacchiotti and Popescu [PP11] call those related words *prototypical words and hashtags*. We refer to both as prototypical keywords indistinctively. For any sensitive issue under consideration, we collect relevant tweets based on prototypical keywords (*e. g.*, *#prochoice*, *#prolife*, *abortion*, *pregnancy interruption*, etc.). Those keywords are extracted from a manually constructed knowledge base of issues and their respective related stances and associated terms.

We build *user documents*, defined as the concatenation of tweets from each user. We represent each user document u as a vector

$$\vec{u} = [w_0, w_1, \ldots, w_n]$$

4 Both are soccer teams from Spain.

where w_i represents the vocabulary word i weighted using TF-IDF [BR11a, Chapter 3]:

$$w_i = \mathrm{freq}(w_i, u) \times \log_2 \frac{|U|}{|u \in U : w_i \in u|}$$

where U is the set of users. Note that the user document can be built with all tweets and retweets for each user, as well as a subset of both. In particular, we consider tweets and retweets, but not replies to other users, as they are less likely to contribute information to the document.[5]

Likewise, for each issue stance we build a stance vector $\vec{s}$, defined as the vectorized representation of tweets containing its prototypical keywords:

$$\vec{s} = [w_0, w_1, \ldots, w_n]$$

with w_i weighted according to TF-IDF with respect to the corpus of user documents.

Using these definitions we can estimate how similar is the language employed by a specific user with the known stances of a specific issue. Formally, we define a user stance with respect to to a given sensitive issue as the feature vector $\vec{u}_s$ containing the similarity of user $\vec{u}$ with each issue stance. In this way, we consolidate all similarities in a *user stance vector*:

$$\vec{u}_s = [f_0, f_1, \ldots, f_{|S|}]$$

where S is the set of stances for the all sensitive issues under consideration, and f_i is the cosine similarity between $\vec{u}$ and the issue stance $\vec{s}_i$:

$$\text{cosine\_similarity}(\vec{u}, \vec{s}_i) = \frac{\vec{u} \cdot \vec{s}_i}{\| \vec{u} \| \| \vec{s}_i \|}$$

Having this representation of user stances, we define the *view gap* with respect to to a sensitive issue between two users as the distance between their respective user stance vectors.

5 Consider the following scenario: user A mentions user B: "*@B hi!*" This tweet indicates that A interacts with B, and it is considered. But a reply to the previous question, like "*@A fine! how are you today?*" and then "*@B fine, thanks!*" do not contribute information.

5.3.3 *Representation of User Interests*

In our context, user interests are important in both parts of the application–as input for the recommender system, but also as part of the data portrait design. In particular, we are interested in evocative text that represents user interests [Don+10]. To find these interests, we focus on the most frequent *hashtags*, *mentions* and *retweets* to other users, *n-grams* (*e. g.*, to convert the two words *New York* into a single keyword, *New_York*), and *links* present in a user document, considering domain name and the first directory in the URL. Our assumption is that users tweet regularly about their interests and they express themselves through the aforementioned ways.

To estimate n-grams, we use an implementation of word collocations [Mik+13] available on the *gensim* software library [ŘS10]. We keep the top-250 most frequent tokens, which, jointly with their corresponding frequencies in the user document, are defined as the list of *user interests* for a given user.

5.3.4 *Recommending People of Opposing Views with Shared Interests*

For each user we have a stance vector that describe her/his position with respect to to given issues, as well a the list of user interests with their respective weight. We approach the problem of recommending people of opposing views, but with shared interests, as a content-based problem. We propose to build an inverted inverted index [BR11a] of tweets from people in a candidate pool for recommendation, allowing user interests to be used as search queries.

Algorithm 5.2 formalizes our algorithm. Its input is the inverted index I, a list of user interests Q and the corresponding stance vectors U for the target user, and the number of desired recommendations. For each user interest, it searches for the highest scored tweet in the candidate pool, with a score estimated as the geometric mean of query relevance and view gap with the author of each candidate tweet. Following this procedure gives a list of recommended tweets where the view gap with others is considered when doing content-based recommendations.

Algorithm 5.2 Recommendation of Tweets from People with Opposing Views.

```
INPUT: I ← inverted index of tweets
INPUT: Q ← set of user interests, sorted by desc. importance
INPUT: U ← set of user stance vectors
INPUT: n ← number of desired recommendations
OUTPUT: R ← set of recommended tweets
  FUNCTION RECOMMEND_TWEETS(I, Q, U, n)
    R ← list()
   FOR ALL q in Q DO
        results ← search(q, I)
   IF empty(results) THEN
            continue
   END IF
   FOR ALL r in results DO
            v ← view_gap(U, r.author.U)
            r.score ← geometric_mean(r.relevance, v)
   END FOR
        R.append(max(results, key = score))
   IF length(results) = n THEN
            break
   END IF
   END FOR
   RETURN R
  END FUNCTION
```

5.3.5 *Data Portrait Design*

In this section we explain the rationale behind our visualization design. To depict user profiles we consider *data portraits* [Don+10], which are *"abstract representations of users' interaction history"* [XD99]. Figure 5.1 displays our design.

The usage of data portraits serves as context to, first, create a self-image for presentation of the target user [Gof59]; and, second, to inject recommendations generated by our algorithm. The rationale behind this idea is that, by using a data portrait, we reinforce non-conflicting interests for users when they browse

Figure 5.1: Our data portrait design, based on a wordcloud and an organic layout of circles. The wordcloud contains characterizing topics and each circle is a tweet about one or more of those topics. Here, the user has clicked on her or his characterizing topic *#d3js* and links to corresponding tweets have been drawn.

their own profiles, allowing to contextualize recommendations according to their interest space.

Profile Construction

The data behind a data portrait of user u is a bipartite graph of two sets of vertices: user interests V_u and relevant tweets V_t. The set V_u is directly built from our methodology. The set V_t is a mixture of two kinds of tweets: those authored or retweeted by u, as well as recommendations generated by Algorithm 5.2.

Although the algorithm specifies how to select the recommended tweets, we still need to define which authored or retweeted tweets are available on V_t. Similarly to the algorithm, for each user interest we select the most popular tweet that matches the interest. As popularity score we consider the number of times the tweet has been retweeted plus the number of times the tweet has been marked as favorite by others.

In the bipartite graph, the set of edges E contains connections between user interests and their corresponding tweets authored/retweeted by the target user, as well as recommendations based on them. Note that it is possible that one interest is linked to many tweets; and one tweet to many interests. In the latter case, we refer to primary interest as the more relevant interest associated to a particular tweet, with relevance estimated by the search engine used to associate interests to tweets.

Depicting User Interests

Even though we are working with a graph, depicting user interests in V_u in text format is important, as text *"provides immediate context and detail"* [Don+10]. Several text visualizations exist [WV08; VWV09], however, they refer to structured text. Our scenario does not consider text structure, and, in fact, in our schema there is no semantic relation between user interests in addition to sharing relevant tweets. Hence, wordclouds are a good choice in this aspect, as they are popular depictions of weighted text used with many purposes on the Web, from navigatonial to participatory expressive visualization [VW08; VWF09]. We exploit this by using wordclouds as devices to encourage interaction with relevant tweets.

Figure 5.1 shows our wordcloud design. Font size encodes the frequency of user interests–a bigger size indicates more frequency. Word positioning and color are random, as in typical wordclouds. Unlike typical wordclouds, Figure 5.1 does not show a tight layout, as expected from the popular and recognizable algorithm Wordle [VWF09]. The reason is that we do not do pixel based collision detection between interests; instead, each interest is positioned according to an expanded bounding box. This expansion is needed to make interests easier to click (or touch in case of touch-screens). Figure 5.2 shows an early prototype that displays the mentioned bounding boxes for illustration purposes.

Depicting Tweets, ReTweets and Recommendations

As observed in Figure 5.1, the nodes from V_t are depicted as circles at the center of the data portrait. We display only tweets and retweets from the portrayed user, as recommendations are injected in the portrait through interaction with the displayed nodes, as shown on Figure 5.4.

Influenced by *organic information design* [Fry00], node position is based on an organic model of pattern of florets from Vogel [Vog79], defined in polar coordinates as:

$$r = c\sqrt{n}$$
$$\theta = i \times \phi$$

Where c is a constant that defines how separated the circles are, ϕ is the *golden ratio* (defined as $(1+\sqrt{5}) \cdot 0.5$), and i is the reverse chronological ordered index of each tweet (the oldest tweet will have index 1). The color of each circle is based on the color assigned to the primary interest of the corresponding tweet. Note that interests do not overlap with the circles. Node side is proportional to popularity, and its maximum size depends on screen size and target platform (*i. e.*, touch screens need larger circles). Ideally, the difference between minimum and maximum node size should not be big, as node size should be used as a feature to give the impression of organic variance instead of explicitly encoding the difference in popularity.

When a circle is clicked, a pop-up balloon appears containing the corresponding tweet with a format that resembles the native format in Twitter. The tweet includes native options such as retweet, reply and mark as favorite. To nudge

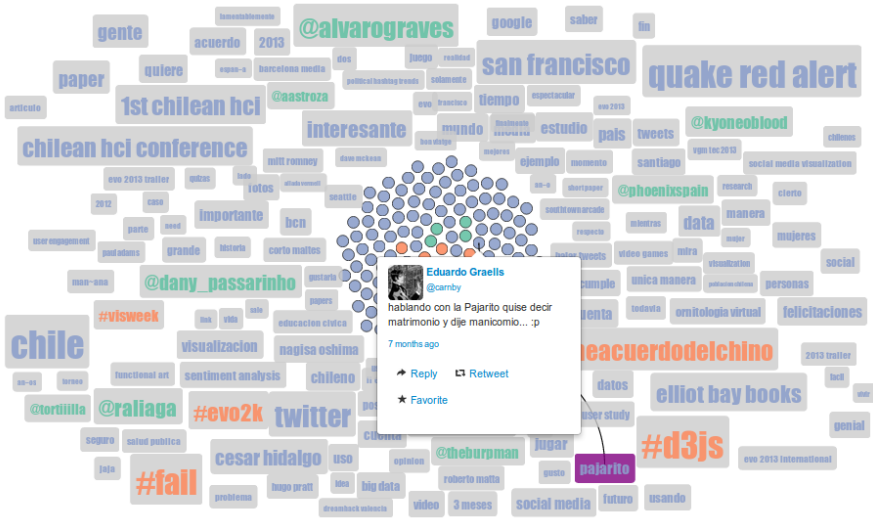


Figure 5.2: Early implementation of the data portrait design. This image displays each interest's bounding box, which are intended to ease clicking.

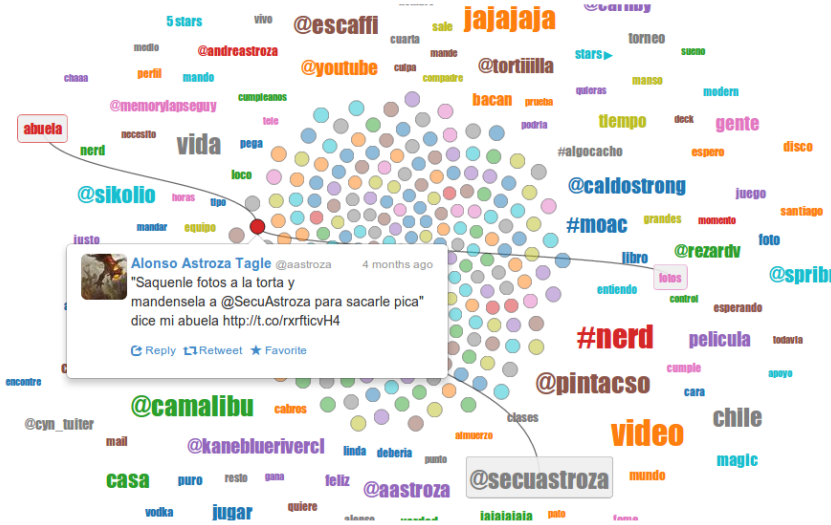


Figure 5.3: State of the data portrait after a circle node has been clicked. A tweet is displayed in a pop-up balloon and links to the corresponding interests are visible.

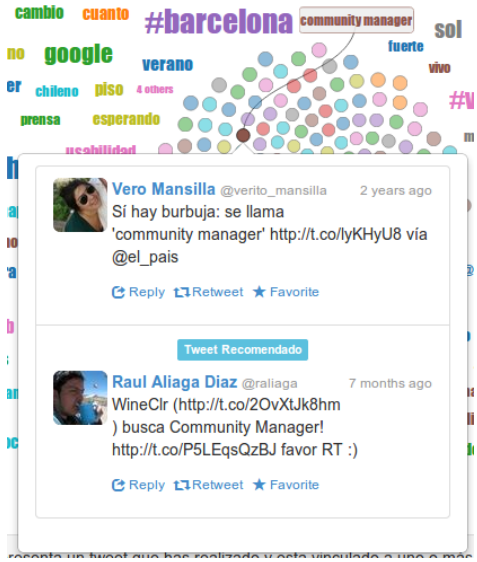


Figure 5.4: Display of tweets inside a pop-up balloon.

interaction with others, in addition to the portrayed individual tweet, we inject a recommended tweet that is related to the primary interest of the current displayed tweet, as shown in Figure 5.4.

A second click on a circle makes the pop-up balloon disappear, allowing the user to have a clear sight of the structure of the corresponding tweet connections. Clicking into empty space on the visualization canvas clears all visible links.

Depicting Links between Interests and Tweets

Contrary to typical graph visualizations, our design does not show all connections between nodes as default, as connectivity in the graph is not the main purpose of the data portrait. We only display links when the user interacts with the data portrait. The following interactions trigger changes in visibility of nodes:

1. Clicking on interests reveals links to the relevant tweets (see Figure 5.1).
2. Clicking on nodes reveals links to the relevant interests (see Figure 5.3).

To keep the organic feel, link paths are generated using *bézier curves*.

Using this data portrait design we expect users to engage in explorations of their interests and find new people to connect with. In the next sections we evaluate this design, as well as the recommendation algorithm, in a formative user study to validate our motivating ideas before incorporating this design into an end-user application.

5.4 CASE STUDY: ABORTION IN CHILE

In this section we describe a case study where we analyze the issue of abortion in Chile using our methodology to estimate user stances.

5.4.1 *Why Abortion in Chile?*

The history of abortion in Chile is long, being declared legal in 1931 and illegal again in 1989. As of 2015, abortion is still illegal, making Chile one of countries with most severe abortion laws in the world [Uni; SB07].

Abortion in Chile as sensitive issue has good properties for analysis, as it is constantly being discussed on the political active population. On one hand, 61% of population was estimated to be catholic, and 21% professed another religion, while only 19% of the population was atheist or agnostic [Pon14]. On the other hand, 63% of the Chilean population was in favor of legalization of abortion in 2013 [CEP13]. The occurrence of several protests around public education, same-sex marriage and abortion, among other sensitive issues, are encouraging the usage of micro-blogging platforms and social networks to spread ideas and generate debates (for a discussion on the student movement in Chile see Barahona *et al.* [Bar+12]). This duality where a majority of population is estimated to have conservative views, but also a majority of population is in favor of legalization of abortion, while a growing portion of the population is asking for reforms using social media as a primary communication and organization device, makes Chile and ideal scenario for analysis.

From a computational point of view, estimating user stances on abortion is simpler (and thus more feasible) than other issues. Even considering the complexity behind the rationale of the different factors that influence a stance on abortion, there are two main stances:

- Pro-Choice: *"emphasizing the right of women to choose whether to abort a pregnancy or to grow it to term"* [Wik15].
- Pro-Life: *"emphasizing the right of the embryo or fetus to gestate and be born"* [Wik15].

Although both stance names have been criticized for being politically framed (*e. g.*, being pro-choice is not being "anti-life"), they are widely recognized by such names, and the discussion of proper names for both stances is outside of the scope of this dissertation.

In spite of the growing discussion in both off and on-line worlds, it is not clear if a real debate exist, or whether the supposed spreading of ideas is just loud broadcasting where no one listens to actually improve their stances, but just to reinforce own beliefs. How to evaluate this is what we propose in this section. In this scenario, homophilic behavior would be to have meaningful discussion only with people from the same abortion stance, with loud broadcasting being to mention people from any stance without engaging in discussion. To confirm or discard these traits in the studied population we applied the methods defined in this chapter. Then, to encourage connecting with others of opposing views in abortion, we created data portraits and inject recommendations of such people having non-conflictive shared interests.

5.4.2 *Dataset Description*

In the context of on-going campaigns for presidential elections, we crawled tweets from July 24th, 2013 to August 29th, 2013 using the *Twitter Streaming API*. Table 5.1 shows a summary of the crawled data. In total, we crawled 367,512 tweets from 57,566 accounts that were geolocated using a gazetteer. Of those tweets, 18,148 are related to abortion, as they contain at least one prototypical keyword (see Table 5.2). The vocabulary size is 38,827, filtering out all keywords that appear in less than 5 tweets.

Initially, we used *query keywords* about known sensitive issues and hashtags: *abortion* (issue), *education* (issue), *same-sex marriage* (issue), *Sebastián Piñera* (president in 2013), *Michelle Bachelet* (candidate), *Evelyn Matthei* (candidate), among others. We searched for keywords about other sensitive issues because, according to our methodology, we will consider the relationship between lan-

Table 5.1: Data crawled from Twitter during July and August 2013.

Data	#
Tweets	367,512
Tweets About Abortion	18,148
Accounts	57,566
Accounts with Abortion-related Tweets	8,794
Vocabulary Size	38,827

guage and user stances. We also added emergent hashtags related to news events that happened during the crawling period. For instance, *#yoabortoel25* is about a protest held on July 25th [Can15]. Figure 5.5 shows the most frequent terms found in our collection. The most prominent words are last names of candidates, namely *Evelyn Matthei*, *Michelle Bachelet*, *Pablo Longueira* and *Laurence Golborne*. The last name of the dictator *Augusto Pinochet* is also prominent. Other prominent keywords are *carabineros* (the police), *censo* (the national level census conducted in 2012, with multiple flaws that were discovered in 2013), *Transantiago* (public transport system in Santiago), *isapres* (the private health system) and *AFP* (the name of the Chilean pension system, composed of several *Administrators of Public Founds*).

We removed tweets in other languages than Spanish, tweets that were not geolocated to Chile according to users' self-reported location, as well as noisy tweets. When crawling tweets about sensitive issues, we noted that abortion related tweets were unlikely to be noisy, *i. e.*, a tweet with abortion related keywords is about abortion as issue. Other issues like public education and the student movement where much more noisier; for instance, hashtags like *#education* are used to promote services by educational institutions. The numbers reported in Table 5.1 consider a cleaned dataset.

Table 5.2: Keywords used to characterize the pro-choice and pro-life stances on abortion. General keywords plus stance keywords were used to find people who talked about abortion in Twitter.

Stance	Seed Tweets	Seed Users	Keywords
Pro-choice	95,173	1,934	#abortolibre, #yoabortoel25, #abortolegal, #yoaborto, #abortoterapeutico, #proaborto, #abortolibreseguro gratuito, #despenalizaciondelaborto, #abortoetico, #abortolegal, #abortosinapellido, #derechoadecidir
Pro-life	10,040	338	#provida, #profamilia, #abortoesviolencia, #noalaborto, #prolife, #sialavida, #dejalolatir, #siempreporlavida, provida, #nuncaaceptaremoselaborto, #chilenoquiereabortos, #conabortonohayvoto, #yoasesinoel25, #somosprovida
General Words	–	–	aborto(s), abortista(s), abortados(as), abortivo(a)... ...(tenses of *to abort* in spanish)
Related Hashtags	–	–	#marchaabortolegal, #bonoaborto, #cifrasaborto, #feminismo
Relevant Accounts	–	–	@elardkoch, @siemprexlavida, @quieronacer, @mileschile, @melisainstitute, @ObservatorioGE
Contingency Words	–	–	terapéutico, violada, violación, violaciones, interrupción, inviabilidad, embarazo, embarazada, feto, embrión, fecundación, antiaborto, feminismo

Figure 5.5: Frequent terms in the collection. Green terms were used as query keywords for crawling.

5.4.3 *Pro-Choice and Pro-Life Stances*

We manually built a list of words, accounts, and hashtags related to abortion and its two stances. We iteratively explored the dataset to find co-occurrences of prototypical keywords like *abortion*, *#abortolibre* (*free abortion*) and *#noal-aborto* (*no to abortion*). Table 5.2 shows the obtained abortion-related terms. For pro-choice and pro-life keywords, the number of seed users and their number of tweets is displayed. These seeds represent whether a user document contained keywords from one stance but not from the other, *e. g.*, a user document that contains at least one pro-choice keyword and no pro-life keywords is considered a pro-choice seed user. As observed in the table, the number of pro-choice seed users outnumbers those of pro-life stance (1,934 pro-choice against 338 pro-life). This does not necessarily indicate the proportion of users from both stances, instead, it might indicate that pro-life users tend to inject content into pro-choice timelines by using their hashtags [Con+11b].

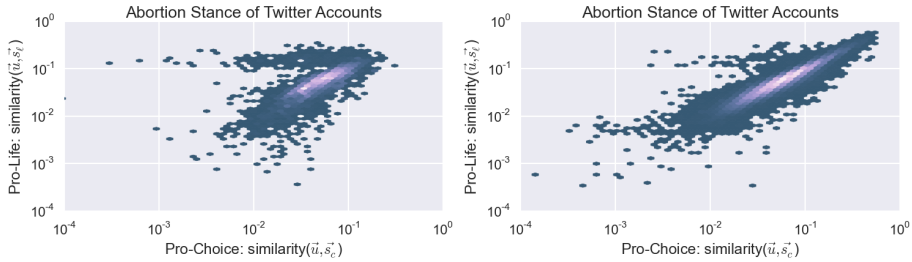


Figure 5.6: Distributions of user stances based on similarity between user vectors and stance vectors (pro-life and pro-choice). Left: stances of users who tweeted about abortion. Right: stances of all users in the dataset.

User Stances

To build the stance vectors of pro-choice and pro-life stances, we concatenated the tweets of the corresponding seed users of each stance. Then, according to our methodology, we estimated the user stances on abortion by computing the cosine similarity between user vectors and the stance vectors. These similarities are displayed with hexagonal binning in Figure 5.6, where the x axis represents similarity with the pro-choice stance vector $\vec{s_c}$; the y axis represents similarity with the pro-life stance vector $\vec{s_\ell}$. We display two charts: one for users who have tweeted about abortion (8,794) on the left, and one that considers all users on the dataset (57,566) on the right. This is possible because the user stance vectors are constructed using all vocabulary employed by seed users; hence, they contain valid weights for words unrelated to abortion, but related to additional issues that those users discussed. Under the assumption that sensitive issues have a degree of correlation among stances in different issues, this allows us to estimate a tendency for all users.

Having estimated similarities with both abortion stances, we define *stance tendency* as:

$$\text{tendency} = \text{cosine\_similarity}(\vec{u}, \vec{s_c}) - \text{cosine\_similarity}(\vec{u}, \vec{s_\ell})$$

We classify users with tendency $\geqslant 0$ as pro-choice, and pro-life otherwise. Similarly, we estimate the *view gap* between two users $\vec{u_1}$ and $\vec{u_2}$ as:

$$\text{view\_gap} = \text{abs}(\text{tendency}(\vec{u_1}) - \text{tendency}(\vec{u_2}))$$

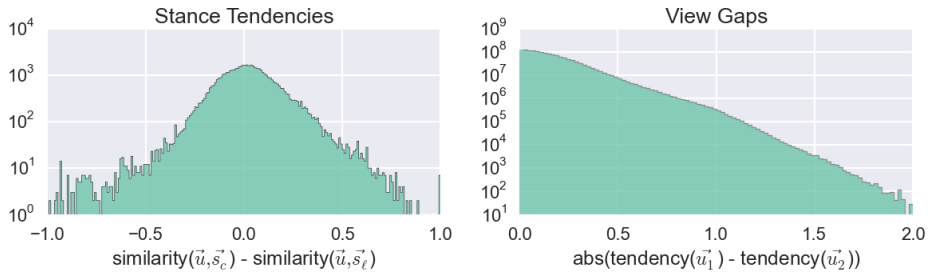


Figure 5.7: Left: Distribution of tendency to pro-life or pro-choice stances for users in our dataset. A positive value means leaning to pro-choice, and a negative value means leaning to pro-life. Right: distribution of *view gaps* in abortion (distances between user tendencies for pairs of users).

Using these definitions of tendencies and view gaps, we are able to estimate how distant two users are in terms of their views on abortion. Figure 5.7 shows the distribution of tendencies (left) and view gaps (right). The median stance tendency is 0.02, showing a slight tendency towards the pro-choice stance: 54.98% of users are classified as pro-choice, while 45.02% of users are classified as pro-life. Pro-choice users published 10.24 tweets in average, while pro-life users published 10.48 tweets in average.

According to CEP [CEP13], 63% of the Chilean population was in favor of legalization of abortion in 2013. Our predicted proportion of user stances does not differ from expectations according to a chi-square test ($\chi^2 = 2.76$, $p = 0.10$). While the Twitter population is not demographically representative of the population, this result indicates that abortion stances are reflected on the micro-blogging platform Twitter.

5.4.4 *Population and Content Distribution*

In Figure 5.8, we explore the properties of all users in terms of their predicted abortion stances. On the left, we consider the complimentary cumulative distribution functions (CCDF) of the number of followers and friends, showcasing similar distributions between stances, as well as to those of Kwak *et al.* [Kwa+10]. The estimated power-law [ABP14] PDF parameters are the same for

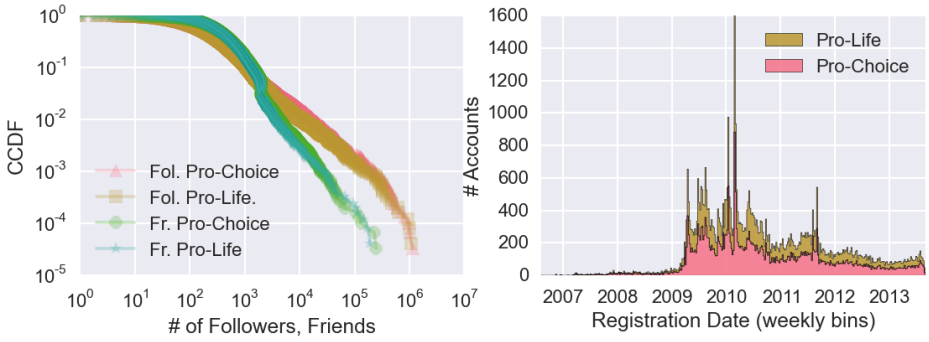


Figure 5.8: Left: Complimentary Cumulative Distribution Function of user connectivity for pro-life and pro-choice users. Right: time of registration of users who tweeted about sensitive issues, according to their abortion stances.

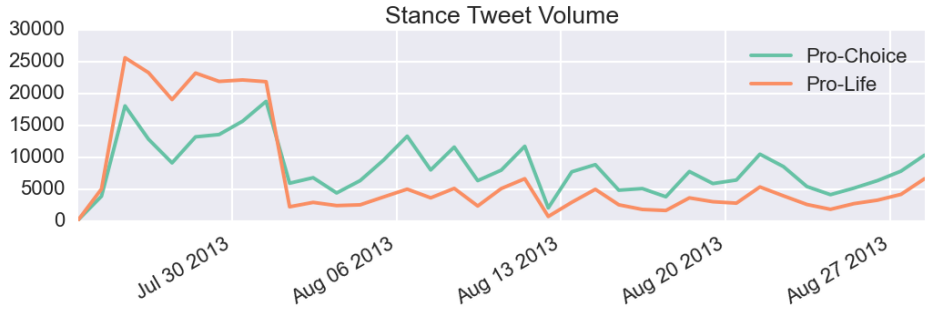


Figure 5.9: Tweet volume per abortion stance.

Figure 5.10: Most associated words with pro-choice (left) and pro-life (right) users according to their self-reported biographies, estimated with *Pointwise Mutual Information* [CH90]. Font size is inversely proportional to PMI rank. Color encodes frequency (the darker, the more frequent).

both stances: $\alpha_{\text{follow}} = 1.19$ and $\alpha_{\text{friends}} = 1.17$. On the right, we show the distribution of registration date of participating user accounts. We see accounts from the beginning of Twitter until the studied period. The distribution is similar to the observed in the previous case study about geographical diversity, a coherent result given that both datasets are focused on political content.

Figure 5.9 shows tweet volume per abortion stance. Between July 24th and August 2th there is a high volume of tweets in comparison with the rest of the dataset. This is related to the national level protest held on July 25th, with hashtag *#yoabortoel25* [Can15]. Surprisingly, being a pro-choice protest, the volume of pro-life tweets surrounding the event date is higher than those of pro-choice users. As mentioned earlier, this might be related with content injection from pro-life users into the pro-choice hashtags.

5.4.5 *Associated Description Words with Abortion Stances*

Having classified users into pro-life or pro-choice, we can explore which words are associated with both stances. In particular, we are interested in the reported *self description* available on each user profile, to see if our stance classification gives coherent result. We measure *Pointwise Mutual Information* [CH90] over the set of vocabulary in self descriptions. PMI is defined as:

$$\text{PMI}(c, w) = \log \frac{p(c, w)}{p(c)p(w)}$$

where c is a stance (*pro-choice* or *pro-life*), and w is a word present in self descriptions. The probabilities can be estimated from the proportions of description in pro-choice and pro-life users, and the corresponding proportions of words. Since PMI overweights words with very small frequencies, we consider only words that appear in at least 50 user descriptions.

Of all users, 54,640 have a valid self-description. Figure 5.10 shows the results of PMI estimation. Some relevant keywords for our context for each stance are (in parenthesis their PMI rank and translation):

- Pro-choice: *#asambleaconstituyente* (#1, a hashtag to support a new constitution), *trabajadores* (#2, *workers*), *progresista* (#3, *progressive*), *feminista* (#5, *feminist*), *mapuche* (#6, indigenous inhabitants of south-central

Chile), *revolución* (#7, *revolution*), *gratuita* (#10, *free*, referring to the student movement and their plea of free education), *ciudadana* (#14, *female citizen*).

- Pro-life: *gremialista* (#1, *guildist*–guildism is a ideology based in Catholic social teachings), *pinochetista* (#2, supporter of Augusto Pinochet), *udi* (#3, abbreviation of *Union Demócrata Independiente*, a right-wing party), *derechista* (#4, *right-winger*), *#matthei2014* (#5, hashtag to support *Evelyn Matthei*, a candidate representing UDI in the presidential elections), *conservador* (#6, *conservative*), *católico* (#13, *catholic*).

As observed, the highest scoring words for the pro-choice stance are related to left-wing prototypical keywords and entities, and the highest scoring words for the pro-life stance are related to catholic and right-wing prototypical keywords and entities. This indicates that our results are coherent with the expected ideology of the two abortion stances considered.

5.4.6 *Homophily in One-Way and Two-Way Interactions*

Having predicted a stance for each user in the dataset, we are able to evaluate if interactions in the dataset are homophilic, *i. e.*, we will test if users tend to interact with people of the same abortion stance. To do so, we study 1-way and 2-way interactions as defined by Quercia, Capra, and Crowcroft [QCC12]. Mentions and retweets are 1-way interactions, where the target user is not necessarily participant of the interaction. When the target users replies to the mention or the retweet, we consider a 2-way interaction. To measure homophily, we estimate the aggregated interactions between users in both stances, and compare their inter-stance proportions with the proportions of predicted stances for all accounts. If interaction behavior is unbiased, then the proportion of interactions between stances should not differ significantly to the proportion of users in each stance.

Table 5.3 shows 1-way interactions between stances. The amount of interactions initiated by pro-choice users is higher than those of pro-life users. Pro-choice users exhibit homophilic behavior: 83.48% of their 1-way interactions are with other pro-choice users, and their proportions of intra-stance interactions differs significantly with the expectations according to a chi-square test,

Table 5.3: 1-Way Interactions Between Abortion Stances. *: $p < 0.001$

Abortion Stance	1-Way	Pro-Choice	Pro-Life	χ^2	w
Pro-Choice	92,507	83.67%	16.33%	33.27*	0.34
Pro-Life	56,320	49.77%	50.23%	1.09	–

Table 5.4: 2-Way Interactions Between Abortion Stances. *: $p < 0.001$

Abortion Stance	2-Way	Pro-Choice	Pro-Life	χ^2	w
Pro-Choice	2,282	78.79%	21.21%	22.91*	0.31
Pro-Life	1,733	27.93%	72.07%	29.55*	0.33

with a moderate effect size ($\chi^2 = 33.27$, $p < 0.001$, Cohen's $w = 0.34$). Pro-life users do not show this behavior–they mention and retweet users as expected given the proportions of users.

Because 1-way interactions fail to capture meaningful discussion, we estimated 2-way interactions, where an interaction between two users exist if they retweet, mention or reply each other in both ways. To avoid bias in the estimation, we only considered each pair (u_1, u_2) once per inter-stance interactions. Table 5.4 shows the number of interactions between stances and their proportions. As observed, the number of 2-way interactions is much lower, and it is similar for each stance (pro-choice: 2,234; pro-life: 2,042). Furthermore, the proportions of interactions with the same stance is similar (pro-choice: 76.45%; pro-life: 74.24%). A chi-square test indicates that both proportions differ significantly from the expectations (pro-life: $\chi^2 = 29.55$, $p < 0.001$, Cohen's $w = 0.33$; pro-choice: $\chi^2 = 22.91$, $p < 0.001$, Cohen's $w = 0.31$), confirming the homophilic behavior in terms of abortion stances in the studied population.

Note that pro-life users show homophilic behavior only in 2-way interactions. The lack of this behavior in 1-way interactions might be related with the content injection we mentioned earlier.

Through this analysis we have observed that it is possible to predict abortion stances for Chilean users participating in political debate on Twitter. In terms of these abortion stances, the user population exhibits homophilic behavior when interacting with others.

5.5 PILOT USABILITY STUDY

Having verified the homophilic behavior of users in our dataset with respect to abortion stances, in this section we evaluate with a user study our data portrait paradigm and user perception of our generated recommendations. Following the procedure from our previous case study, we put emphasis on individual differences related to the context of the problem we are trying to solve, *i. e.*, we observe if *having tweeted about abortion before the study* is an indicator of differentiated behavior.

5.5.1 *Experimental Setup*

Participants

Participants were recruited from social networks using an open call to volunteer in a user study about data portraits. No compensation was offered. We recruited 36 participants.[6] Of them, 25 are male and 11 are female. In terms of age, four were 18–25 years, 14 were 25-30 years, 17 were 31-40, and one was 41-50. All participants were Chileans: 22 from Santiago, four from other locations in Chile, and 10 outside of Chile. When participants were asked to rate their experience with social networks they scored themselves 3.83 (std. 0.81) in average, using a Likert scale from 1 to 5.

The open call did not disclose that the study was about the injection of recommendations from people of opposing views. In a post-study survey we disclosed this information, and we asked participants if they had tweeted about abortion before (20 answered yes, 16 answered no).

6 We discarded one user from a previous report of this study in [GLQ13]. After an inspection of interaction data, we found that this user did not follow the instructions of visiting the site at least three times.

Recommended Tweets and Candidates

Of all Chileans who published tweets in the case study, we selected a group of 4,077 candidates for recommendation. In particular, we considered users that were likely to be regular people, *i. e.*, those who follow less than 2,000 accounts and are followed than less than 2,000. This filtering was made because regular people is arguably more prone to discuss their own interests, unlike popular accounts which may be from media outlets, blogs, or celebrities. From those regular users, we crawled 1,400,582 tweets from December 6th, 2013 until January 3th, 2014. To have an index for recommendation, using the *gensim* [ŘS10] library we built an inverted index of tweets. For each participant we estimated user interests and queried the index using these interests as queries. Since we had estimated each candidate stance on abortion before, this estimation was used to calculate the view gap between the authors of search results and the target participant, allowing us to rank tweets according to our methodology.

Apparatus

Participants were tested in a on-line setting. Our data portrait design, as well as the baseline (see Figure 5.11), were implemented in HTML and Javascript using the *d3.js* library [BOH11]. Before the start of the experiment, we explained to participants that we have built a visually explorable characterization of them and that we will display related tweets to their characterization. Each participant was given a unique URL with their portrait to visit.

After the experiment, participants filled a post-study survey with two parts: one part contained usability questions, and a second part with questions related to the sensitive issue of our case study. On the first part, participants answered the following questions using a Likert scale from one to five:[7]

1. How much did you enjoy using the application?
2. Would you use the application if it was integrated in Twitter?
3. How much did you feel represented by the portrait?
4. Do you think the portrait allows you to discover patterns in your behavior?
5. How similar were recommended tweets to your tweets?

7 All questions were asked in Spanish. We translate them in English for ease of understanding.

Figure 5.11: Baseline interface of the data portrait pilot study. On the top, we use a standard wordcloud to display *user interests*. On the bottom, two timelines are displayed using a typical Twitter format. The timeline on the left displays *user tweets*. The timeline on the right displays *recommended tweets*.

6. How interesting were the recommended tweets for you?
7. How serendipitous were the recommendations?[8].

Participants were not told that the recommendations they would receive were from people with eventually opposing views, and they were not told about the sensitive issues aspect of the experiment until the second part of the post-study survey. There, we explained why we asked "*Have you tweeted about abortion before the study?*" in the first part of the post-study survey.

Design and Conditions

The experiment used a *between-groups* design. We defined three conditions:

1. *Baseline* (N = 12). It displays user interests using a standard wordcloud. Participant tweets and recommended tweets were displayed using parallel lists, formatted in a similar way to the standard timeline in Twitter (see Figure 5.11). Recommendations were generated considering query relevance only (*i. e.*, excluding view gaps).
2. *Condition I* (N = 12). It uses the baseline design, but recommendations were generated using our recommendation algorithm (*i. e.*, considering view gaps and relevance).
3. *Condition II* (N = 12). User interests and tweets are visualized using our data portrait design, and recommendations were generated using our algorithm.

Participants were assigned randomly to each condition.

Task

We asked participants to visit their portrait during three consecutive days, and to browse their portraits for as long as they want, but for a minimum of three minutes. If participants tweeted during the days of experiment, their portraits were updated. They were encouraged to explore their user preferences, but we did not explicitly encourage them to follow others.

8 Note that "*serendipity*" was translated from *"surprising"*, as *serendipity* does not exist in Spanish. Every other relevant word in our context has been directly translated.

Figure 5.12: Distributions of user variables from the pilot user study, plotted using violin plots [HN98].

5.5.2 *Quantitative Results*

Figure 5.12 displays the distributions of survey answers for all Likert scale questions. Over these results we performed a factorial ANOVA, which allows us to determine if there is a statistical dominance of at least one group when comparing means of multiple groups. The model assumptions of ANOVA are: normal distribution of the data, homoscedasticity of each group (equal variances) and independence of observations. Our data is ordinal and, according to Shapiro-Wilk tests, is not normal. Figure 5.12 showcases the distributions of responses from the post-survey by using *violin plots* [HN98]. However, ANOVA is robust to normality violations [Sch+10] and has been commonly applied to Likert-scale data in human computer interaction literature (*e.g.* 80.6% of `CHI2009` papers that analyzed Likert response data used parametric methods [KNM10]) when the variances in groups are equal. We tested our dataset using Levene's test for equality of variances for all groups, and found that all groups have equal variances; additionally, our analysis considers a heteroscedasticity-corrected coefficient covariance matrix when performing ANOVA.

We first evaluated the following factorial model:

$$Y = C(\text{ui}) \times C(\text{recommendation}) \times C(\text{posted\_about\_abortion})$$

Where $C(IV)$ creates dummy variables for the corresponding categories of the independent variable IV, and multiplications represent both additions and interactions between factors. If the interactions were found to be not significant, then we ran a new ANOVA without interactions for the following model:

$$Y = C(\text{ui}) + C(\text{recommendation}) + C(\text{posted\_about\_abortion})$$

If results were significant for a variable, we report the dominant group and its F-value, and then perform a robust linear regression to estimate the significance of each factor.

Enjoyment

When replying *"How much did you enjoy using the application?"*, the mean value reported by all users is 3.78 (std. 0.83).

There is significant interaction between recommendation condition and *having abortion-related tweets* ($F(1, 30) = 7.13,\ p = 0.01$). When users receive non-opposing recommendations (the *regular* condition), and have published tweets about abortion, then the *enjoyment* of the application is lesser than when not (linear regression $\beta = -0.5$, $p < 0.001$).

Would Use the Visualization

When replying *"Would you use the application if it was integrated in Twitter?"*, the mean value reported by all users is 3.34 (std. 0.84). No differences were found between conditions.

Identification

When replying *"How much did you feel represented by the portrait?"*, the mean value reported by all users is 3.50 (std. 0.77).

There is a significant effect of recommendation condition ($F(1, 32) = 4.79$, $p = 0.04$). When users receive non-opposing recommendations their identification with the portrait increases (linear regression $\beta = 1$, $p < 0.001$).

Discover Patterns

When replying *"Do you think the portrait allows you to discover patterns in your behavior?"*, the mean value reported by all users is 3.75 (std. 1.08). No differences were found between conditions.

Recommendation Similarity

When asked to rate *"Recommendation Similarity"*, the mean value reported by all users is 2.53 (std. 0.91).

There is a significant effect of having abortion-related tweets ($F(1, 32) = 9.86$, $p < 0.001$). When users published abortion-related tweets before the study, their perception of similarity of recommendations increases ($\beta = 1$, $p < 0.001$).

Recommendation Interestingness

When asked to rate *"Recommendation Interestingness"*, the mean value reported by all users is 2.69 (std. 1.14).

There is a significant effect of having abortion-related tweets ($F(1, 32) = 13.29$, $p < 0.001$) When users published abortion-related tweets before the study, their perception of similarity of recommendations increases ($\beta = 1.236$, $p < 0.001$).

Recomendation Serendipity

When asked to rate *"Recommendation Serendipity"*, the mean value reported by all users is 2.97 (std. 0.97). No differences were found between conditions.

5.5.3 *Qualitative Analysis*

We included open questions in the post-study survey to understand user views about our proposed paradigm. When quoting user feedback, we use Pi to refer to participant i.[9]

9 The answers to the open questions have been translated from Spanish to English.

User Interface and Data Portrait Design

In the answers to the question *"How would you describe the application you have used?"*, the high rating obtained with respect to enjoyment is reflected in the participant responses :

> "I liked the way in which you select the points when you click on a word. I also liked a lot the colors and the tag cloud" [P9, Cond. II].
>
> "Didactic, fun and colorful" [P13, Cond. II].
>
> "Friendly. Clear in terms of concepts and visual representation of the information" [P18, Cond. II],
>
> "I like the connections between tweets based on keywords. It is useful for people that curates their content. I also liked the relations with other users" [P24, Cond. II].
>
> "It is a novel idea. At the beginning I did not understand how it worked, but after a couple of clicks I managed to find the 'rhythm"' [P28, Cond. II].

Responses to the question *"What would you change or add to the application you have used?"* contained several suggestions. Some participants wrote that the proposed design could be improved by considering time in user interests and visualization:

> "I would add the option to filter by time. For instance, to visualize the same things one year ago, two years ago, etc" [P5, Cond. II].
>
> "It should be more up to date, as it contained some old information" [P13, Cond. II].
>
> "There were extremely old tweets (+4 years). This devalues the application, because in general my usage of Twitter is 'now'. I understand the need to see older tweets, so it would be good to have a time filter or a way to narrow the time window" [P28, Cond. II].

Other users mentioned typography and colors:

> "I would change the term colors. Intuitively, I try to create relations between concepts of the same color" [P4, Cond. I].
>
> "Color should mean something, like a common category" [P19, Baseline].
>
> "The typography (*Impact*, I think) is not very appealing. In addition, the colors seem to be default choices" [P25, Cond. I].

Finally, other users would like to see a personalized profile:

> "It should have customizable backgrounds" [P10, Cond. I].
>
> "I would like to customize the background image" [P35, Cond. I].

This indicates that, even though the content is personalized, some elements of the user interface should be personalized also.

With respect to interacting with their data portraits, some participants stated they were not interested in doing so:

> "Interesting, but not dynamic enough, too static to take a real benefit from it" [P3, Baseline].
>
> "An interesting 'gimmick' but not necessarily useful for the typical user I know from Twitter" [P10, Cond. I],
>
> "Pretty, but not so useful, of superfluous navigation" [P17, Cond. II].

This was expected, as *Casual InfoVis* systems [PSM07] are not there to solve a task, and as such can be considered as not very useful. Recall that our aim was to define a paradigm with no task in mind, apart of just "browsing".

Discoveries and Wordclouds

In the post-study survey we included the following item: "*Here you can tell us if you discovered something about your profile, if you found what you expected or not, or even something that might have surprised you. You can also write about what was wrong with the content of your profile, and suggest how we can improve it*". As in previous work [VGD06], many users discovered something about themselves:

> "The cloud shows many curious terms that sometimes you do not notice how frequently you use them" [P2, Baseline].
>
> "I did not know that I wish so many happy birthdays in Twitter" [P11, Baseline].
>
> "I found some tweets I did not even remember I had written" [P14, Cond. I].
>
> "I expected many things in terms of words and concepts, but at the same time I found novel things, recurrences that I would have never thought I make when I tweet" [P18, Cond. II].
>
> "I was surprised by the most highlighted concept, it was a discovery. I knew it was important but not that it was the most. Really good finding" [P22, Baseline].
>
> "I was surprised by the amount of tweets associated to certain concepts I did not consider I was using them so much, but here they were exposed. I liked it because it helps you to understand your profile" [P23, treat. II].

But not all feedback was positive in this regard. Some users were distracted by the amount of user interests detected:

> "Maybe [the wordcloud] could be refined and show less concepts. Too much things were flying on the screen" [P20, Baseline].
>
> "There could be a better criteria to show something on the cloud, because I saw meaningless common words" [P30, Cond. I].

Participant 20 refers to the initial animation of the wordcloud, implemented using a force directed graph with *d3.js* [BOH11].

Recommendations

Even though our recommendation obtained regular scores, some participants had good things to say about the recommendations. Some even explicitly mentioned similarity, both political and in shared interests. Note that these answers were written by participants before knowing the political aspect of the experiment:

> "Effectively, I discovered someone new. I did not follow them, but that is another thing, in general I do not follow many people because I want to keep my timeline clean. But I did consider following new people…" [P28, Cond. II].
>
> "I followed a couple of users with similar social/political opinions" [P32, Cond. I].
>
> "I was surprised to see a Twitter user related to the majority of my concepts in the application. Even though we were not tweeting about similar stuff, the content of his/her tweets was interesting for me. I mean, if this would have been part of Twitter, I would have followed this person" [P34, Cond. I]
>
> "I have followed only those who are similar to what I have tweeted about…" [P35, Cond. I].
>
> "I followed some, because they seemed to be intelligent and had political opinions similar to mine. Also, when I clicked words related with music, I wanted to follow people with the same musical tastes" [P36, Cond. I].

These kinds of discoveries are coherent with the literature. As shown by Chen *et al.* [Che+09], content-based recommendations are better for discovery, in contrast to network-based recommendations which are well (and better) received. Moreover, it is interesting to note that some users "considered following" but at the end did not follow. In this aspect, Brzozowski and Romero [BR11b] found that organic network growth is different to follow behavior when users were asked users to consider following others. Yet, the fact that the intention appeared is a good result, given the scope of the study.

Some users explained why they just had intentions, or why they did not consider following at all:

> "In general I don't follow people in spontaneous ways, but when friends recommend them or when I read them on retweets. Maybe if I continue using the application I would eventually follow someone" [P1, Baseline].

> "I didn't find the recommendations interesting, maybe because I tend to ignore Twitter recommendations, even on the official page. My main source of recommendations are my friends. Personally, I don't look for recommendations, even when they are directly related to my tweets" [P10, Cond. I].
>
> "Only on 10% of the time I consider recommendations on any social network. Also, in 70% of cases I did not see a relation between the recommendation and the concept. Maybe our profiles were similar, but I didn't realize that, nor I was going to see his/her profile. I'm not one of those who follow everyone on Twitter" [P20, Cond. I].
>
> "In general I like to mark as favorite, because it allows me to have a timeline of funny tweets. I almost never reply or retweet" [P25, Cond. I].
>
> "I didn't follow anyone, in general I don't follow too much people because I like to keep my timeline clean. But I did consider following recommendations" [P28, Cond. II].

When recommendations were not well received, one reason was the vector space model search strategy. As implemented in our prototype, it did not distinguish between meanings:

> "In general, recommendations were pretty random, based only on a couple of words and not on the general theme of my tweet" [P9, Cond. II].
>
> "Recommendations were similar syntactically but not semantically [...]. They were interesting, in the sense of seeing how others use the same words in different contexts, but because of that the other users were not topically relevant for me. It was surprising, but not in the sense of *'ah, this person is writing about the same things'*" [P18, Cond. II].
>
> "I had the feeling that precision in recommendation was greater for central terms. They become more imprecise for the rest." [P26, Cond. I].

> "For instance, looking at a recommendation for a tweet where I mentioned *Harry Potter and the Philosopher's Stone*, I got a recommendation about *The Bible*" [P30, Cond. I].

However, this problem is a limitation of the vector space model, and not from our core idea of considering opposing views.

5.6 SUMMARY OF CASE STUDY AND PILOT STUDY RESULTS

Abortion as a Sensitive Issue

Chile is a highly polarized country on a number of sensitive issues, in particular in abortion, and this polarization is reflected on the proportion of abortion stances predicted by our methodology. This results is aligned with our motivation. We confirmed homophilic behavior in terms of abortion stances, and our classifier gave qualitatively coherent stances for users, according to the analysis based on Pointwise Mutual Information of self descriptions by users. In this aspect, our methodology obtained good results when modeling user behavior in terms of political discussion in Chile.

Pilot Study and User Feedback

Qualitatively, user interests were well received, as they allowed users to discover new things. Our design was well received also, but there are many issues to be considered in a future re-design, namely: consideration of *time*, a meaningful *color palette for words*, *personalization* of the data portrait, and better *readability*. We consider these items in the next section. Quantitatively, no differences were found with respect to design.

Presence of Abortion-related Content

Likewise our previous case study from Chapter 4, individual differences matter on how users perceive information, as participants who have tweeted about abortion evaluated recommendations as more similar to them, and more interesting, than users who have not tweeted about abortion. Users interested in politics are more vocal [CR13b], and, according to our results, they also seem to be more receptive of recommendations.

The effect of using recommendations considering opposing-views in abortion had a measurable negative effect on self-identification with the data portrait: those who received opposing recommendations experienced lesser identification. This is coherent with our motivation based on group polarization, where users who interact with like-minded others reinforce their beliefs and identity [ML75; Sun09].

We found a statistical interaction between having tweeted about abortion and the recommendation condition: participants who have tweeted about abortion and received non-opposing recommendations experienced less enjoyment with the application. We believe this result can be explained by the flaws in the search strategy used to rank recommendations. Because users with abortion-related tweets are more receptive to receiving recommendations, in the case where recommendations where non-opposing the only scoring factor was query relevance in terms of user interests. As observed on the qualitative results, the relevance of search results (and thus, recommendations) was below expectations of users, who are used to systems with powerful capabilities, such as *Who to Follow* at Twitter [Gup+13].

In summary, these results support the motivations behind our application, but there are problems (in visual and algorithmic design) that need to be fixed if we want to encourage users to interact with recommendations. We do this in the next sections.

5.7 INTERMEDIARY TOPICS

In this section we introduce the *intermediary topics* concept, as a way to surpass the limitation introduced by using the vector space model when recommending tweets based on user interests.

As reported by users, the main reason behind the low quality of our content-based recommendations is the lack of meaning by querying user interests. Since meaning is shaped by context, we propose to consider context based in word co-occurrences to remove this limitation from our methodology. To do so, we use *Latent Dirichlet Allocation* [BNJ03], a generative topic model that clusters words based on their co-occurrences in documents, and defines latent topics

that contribute words to documents. In previous work, Ramage, Dumais, and Liebling [RDL10] have used LDA to model micro-blogs; in the same way we create user documents. Hence, in this section we define how to estimate which topics, from the set of latent topics generated by LDA, are suitable to be used for recommendation of people of opposing views. We define those as *intermediary topics*. We demonstrate that it is possible to find and quantify these intermediary topics, and then we define a new algorithm to recommend people of opposing views, having intermediary topics as input.

5.7.1 *Topic Graph and Information Centrality*

We explore the topical diversity of *user documents* by estimating latent topics with LDA. Then, we build an undirected graph where each latent topic is a vertex, and two vertices are connected if both corresponding topics contribute to the same document. We build an undirected *topic graph* where each LDA topic is a node, two nodes are connected if the two corresponding topics contribute to the same document, and edges are weighted based on the fraction of documents that contributed to it.

On the topic graph, we compute *current flow closeness centrality* [BF05] of nodes, which is equivalent to *information centrality* [SZ89]. For each node, it is defined as the inverse of the average of distances with the least resistance (as in an electrical network) to other nodes. By considering resistance, we expect to measure the degree a topic might represent a shared non-challenging interest between two users.

Finally, we define *intermediary topics* as topics whose centrality is higher than the median of centrality of the entire graph.

5.7.2 *Chilean Users and Intermediary Topics*

As a continuation of our case study, we analyze the same set of candidates used for recommendation in the previous section. Recall that this set is composed of 4,077 users who published 1,400,582 tweets. Those tweets are generic, *i. e.*, are those available on their accounts, and have not been crawled using key-

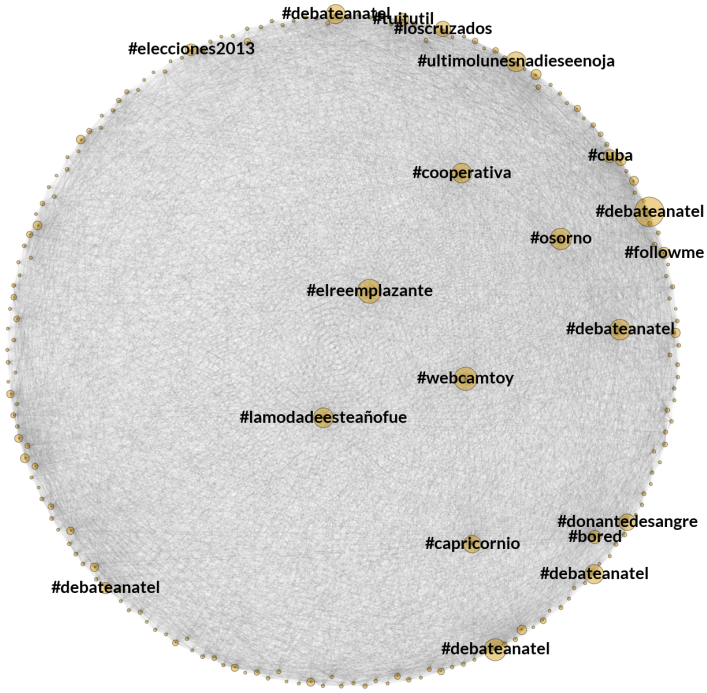


Figure 5.13: Topic Graph. Node size is proportional to centrality. The top decile of central nodes has been labeled with their most contributing hashtags.

words. Jointly with our abortion stance estimation of those users, this makes this dataset useful to test our intermediary topics proposal.

Topic Graph

We ran LDA with $k = 200$, built the topic graph and estimated information centrality as defined by our methodology. After removing junk topics which do not contribute to any user document, the graph contains 198 nodes and 6,906 edges. The median centrality is 1.23×10^{-4}, and its maximum value is 1.64×10^{-4}. The graph is visualized using a spring-based layout on Figure 5.13, where the top decile of central nodes is labeled with the most contributing hashtag to the corresponding topic. As noted on the chart, the graph is dense, because connections between almost all topics exist.

We analyze three variables and their relation with centrality, as well as their differences between intermediary and non intermediary topics: the percent of users that each topic contributes to (Figure 5.14 Left); the probability of abortion keywords to contribute to each topic (Figure 5.14 Right), estimated using the LDA model; and the stance diversity (Figure 5.14 Center), which is the *Shannon entropy* [Jos06] with respect to the predicted abortion stances for all users related to a topic:

$$\text{diversity} = \frac{-\sum_{i=1}^{|S|} p_i \ln p_i}{\ln |S|}$$

Where S is the set of stances, and p_i is the probability of stance i, estimated from the fraction of users assigned to each stance according to our methodology.

Proportion of Users

We observe that central topics have much more users than non-central ones: as users increment, centrality does. This is confirmed by a Spearman ρ rank-correlation of 0.99 ($p < 0.001$) between proportion of users and centrality. The maximum proportion of users a topic contributes to is 78.78%, the median value is 0.56% and the mean is 4.13%. The mean for intermediary topics is 7.99%, and for non-intermediary topics, 0.26%. This difference is significant according to a Mann-Whitney U test ($U = 12.10$, $p < 0.001$). Hence, intermediary topics are more populated than non intermediary topics. This is an expected result, because topic graph construction is based on how topics are related to users.

Stance Diversity

An interesting property is seen on stance diversity, which is very low or very high–there are no intermediate nodes in this aspect. Nodes with high stance diversity can have low centrality, but they concentrate in the upper middle of the chart. The maximum diversity of a topic is 1, its median value is 0.97 and its mean is 0.91. The mean for intermediary topics is 0.96, and for non-intermediary topics, 0.86. This difference is significant according to a Mann-Whitney U test ($U = 3.30$, $p < 0.001$), meaning that intermediary topics are more likely to contain a greater diversity of people with different views on abortion than non intermediary topics.

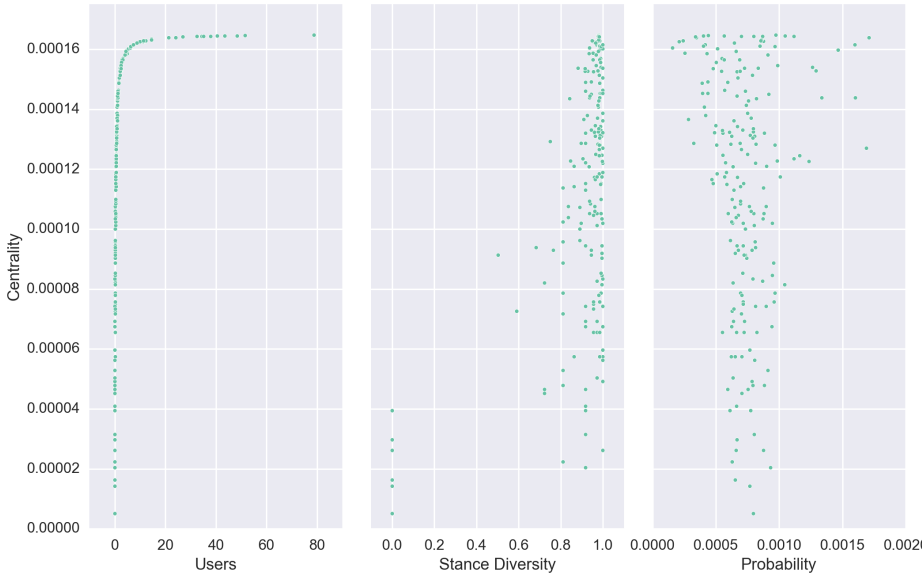


Figure 5.14: Relationship between topic information centrality [BF05] and the percent of users the topic contributes to (left), the abortion-stance diversity estimated with *Shannon entropy* [Jos06] (center), and the probability of abortion-related keywords to contribute to each topic (right).

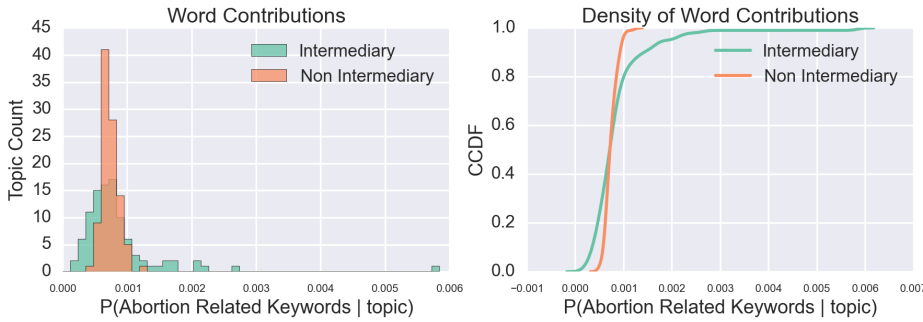


Figure 5.15: Left: Histograms of abortion-related keywords contributions to intermediary and non-intermediary topics. Right: Cumulative Density Function .

Topical Probability of Abortion-Related Vocabulary

Using our set of prototypical keywords, we can estimate the probability of abortion-related vocabulary to contribute to specific topics $P(A \mid t)$, where A is the set of keywords, and t is the target topic:

$$P(A \mid t) = \sum_{i=1}^{|A|} P(w_i \mid t)$$

Where w_i is the ith word in A. Note the LDA model allows us to estimate $P(w_i \mid t)$ directly.

Figure 5.15 displays the distributions and CCDFs of probabilities for intermediary and non intermediary topics. Although the distributions hint a potential difference, given the shape of the curves, this difference is not significant according to a Mann-Whitney U test ($U = -0.59$, $p = 0.55$).

Table 5.5 shows the top-10 intermediary topics in terms of amount of users, topical probability of abortion-related vocabulary, most contributing hashtags and most contributing mentions. All topics in the table have maximum centrality, as well as diversity 0.98. We observe several hashtags and accounts that validate our motivation of finding non challenging information in these topics. For instance, the following hashtags are about non challenging themes: *#minombrees* (topic #1, music TV program), *#elreemplazante* (topic #2, TV program), *#loscruzados* (topic #3, soccer team), *#parapasarlaspenasyo* (topic #5, means "to forget sad moments I..."), etc. However, political content is present, like the hashtag *#debateanatel*, that refers to a TV debate between presidential candidates. Considering the nature of this event, it can be assumed that people from many political stances participated in the discussion.

Regarding mentions, the *@youtube* account is present in many topics, as well as media outlets (like *@biobio* and *@el_dinamo*). Other non-confronting accounts are *@FIL0S0FIA* (light tweets about philosophy), *@horoscoponegro* (astrology), *@dondatos* (a retweet network of services), *@derechosdigital* (a NGO about digital rights) and *@Rh_Negativos* (a supporting group for people with *RH-* blood type). Likewise with hashtags, political accounts of presidential candidates do appear, like *@evelynmatthei* and *@comandomichelle*. However, this could happen because people from all political stances mention them.

Table 5.5: Top-10 Central Latent Topics with Top-5 Contributing Hashtags and Mentions.

Users	Probability	Hashtags	Mentions
78.78%	0.000973	#debateanatel, #bachelet, #chile, #minombrees, #elreemplazante	@biobio, @youtube, @Cooperativa, @el_dinamo, @thecliniccl
51.51%	0.000440	#elreemplazante, #elinternadomega, #debateanatel, #meviolentamuchoque, #vamosbulla	@puccabell, @youtube, @CaataNavarroN, @don_bestian, @FIL0S0FIA
48.37%	0.000401	#webcamtoy, #loscruzados, #lamodadeesteañofue, #selfie, #wweenlared	@youtube, @FIL0S0FIA, @Cooperativa, @horoscoponegro, @GuatonParriyero
43.49%	0.001048	#debateanatel, #votoevelyn, #bachelet, #sisepuede, #minombrees	@nicolaslopez, @sylvi_r, @marciavargass, @Ceci1222, @melnicksergio
40.59%	0.000567	#osorno, #losangelescl, #parapasarlaspenasyo, #elecciones2013, #cambiaeltitulodeunapeliculaporchurrasco	@youtube, @LaMalvada_, @jalyliento, @dondatos, @biobio
37.90%	0.002261	#debateanatel, #yovotomichelle, #costaneranorte, #michelle, #raportero	@biobio, @youtube, @PrensaMichelle, @comandomichelle, @rne_margamarga
35.10%	0.000801	#lamodadeesteañofue, #np, #virgo, #mesiguestesigoentrechilenosque, #parapasarlaspenasyo	@youtube, @puccabell, @jalyliento, @Karol_LuceroV, @_DeJavierTo
34.36%	0.000871	#cooperativa, #cnnchile, #noticias, #debateanatel, #donantedesangre	@biobio, @thecliniccl, @dondatos, @el_dinamo, @Cooperativa
32.70%	0.000801	#debateanatel, #yovotomichelle, #paisaje, #kawaii, #elmatinaldechile	@youtube, @davidpalma77, @derechosdigital, @biobio, @TherealLarrymoe
35.07%	0.000701	#ultimolunesnadieseenoja, #debateanatel, #noticias, #votoevelyn, #sisepuede	@Rh_Negativos, @biobio, @Cooperativa, @latercera, @evelynmatthei

5.7.3 *Recommending People with Intermediary Topics*

Through an extension of the initial case study about abortion in Chile, we have confirmed that intermediary topics do exist and are measurable. The existence of these topics allow us to improve our methodology to recommend people of opposing views. In particular, the usage of LDA allows us to surpass our previous limitation where words appeared in recommended tweets, but they lacked the original meaning of the specific user interest.

Generalization to Political Scenarios

A question that arises regarding intermediary topics is: *does the definition of intermediary topics hold when considering general political views instead of a specific sensitive issue?* We propose that it does because, by definition, intermediary topics only need the estimation of information centrality [BF05; SZ89]. Even though intermediary topics do not need *political view gaps* as input, we found that they have a diverse population in comparison to non intermediary ones (with significant differences according to a Mann-Whitney U test). This means that intermediary topics are a good proxy to include people of opposing views in recommendations.

Rationale

Our proposed approach is based on recommending people of shared interests. Given that we will use intermediary topics to generate recommendations, and that a tweet is too short to be reliably modeled by LDA, we need to recommend user accounts instead, which provide reliable *user documents* to model with LDA [RDL10]. To recommend people, we estimate a content-based distance by using the LDA model, and then estimate user similarity based on intermediary topics. For instance, this would allow the system to select a candidate that has intermediary topics with the target user, but it is distant in terms of political content if the target user is politically vocal. In an opposite case, where users are similar in political content but not on intermediary topics, the candidate would not be selected given that there are no shared interests. Hence, likewise our previous algorithm, we treat recommendations as a content-based ranking problem.

Candidate Scoring

Each candidate for recommendation is scored using a *F-Score* [BR11a] of *latent topical distance* and *similarity with respect to intermediary topics*:

$$\text{score}(u_1, u_2, \gamma) = (1+\gamma^2) \times \frac{\text{similarity}(u_1, u_2) \times (1 - \text{distance}(u_1, u_2))}{\gamma^2 \times (1 - \text{distance}(u_1, u_2)) + \text{similarity}(u_1, u_2)}$$

Where the coefficient γ indicates the importance given to distance in comparison to the importance given to similarity. For instance, $\gamma = 1$ gives equal importance to both factors, $\gamma = 0.5$ gives more importance to distance, and $\gamma = 2.0$ gives more importance to similarity.

We describe how to estimate distance and similarity next.

User Features

Before estimating distances and similarities, we need to define a feature vector for any user u:

$$\vec{u} = [p_0(u), p_1(u), \ldots, p_k(u)]$$

Where k is the number of latent topics, and $p_i(u)$ is $P(t_i \mid u)$ as defined by the LDA model for a topic t_i.

Latent Topical Distance for Users

Given two users, u_1 and u_2, we define their topical distance as the normalized *Kullback-Leibler Symmetric Distance*, defined by Bigi [Big03] as:

$$\text{KLD}(u_1 \parallel u_2) = \sum_{i=0}^{k} \{\vec{u}_1[i] - \vec{u}_2[i]\} \log \frac{\vec{u}_1[i]}{\vec{u}_2[i]}$$

To normalize a distance into the range $[0, 1]$, given a set of distances, we divide each one by the maximum distance found.

Similarity Considering Intermediary Topics

Although topical distance is a strong indicator of similarity, we estimate *similarity with respect to intermediary topics* because we want to be able to recommend

users who might not be close in distance terms (which could happen because of ideological differences), but that may share intermediary topics. The set of intermediary topics for user u is defined as:

$$IT(u) = \{i : \vec{u}[i] \geqslant \varepsilon,\ t_i \text{ is intermediary topic }\}$$

Where ε is a threshold for topical significance, which depends on the context. For instance, the default value used in the *gensim* library is 0.01 [ŘS10].

We define similarity with respect to intermediary topics as the *Jaccard Similarity* between two users:

$$JIT(u_1, u_2) = \frac{|IT(u_1) \cup IT(u_2)|}{|IT(u_1) \cap IT(u_2|}$$

Using this formula, when two users share all intermediary topics, $J(u_1, u_2) = 1$, and when users do not share any intermediary topic, $J(u_1, u_2) = 0$.

Algorithm Formalization

Having estimated a measure of how close two users are, as well as how similar their sets of intermediary topics are, we can formalize our algorithm to recommend people with intermediary topics.

The algorithm is described as follows: given a target user u, a candidate set of recommendations C, a balancing factor γ, and the number of desired recommendations n, estimate the defined score for all candidates and return the top-n scored candidates. The algorithm is described in pseudo-code in Algorithm 5.3.

In this section we introduced a new algorithm to recommend people of opposing views, based on our definition of intermediary topics. In contrast with our previous recommender algorithm, this time we recommend people instead of tweets. This implies that the depiction of recommendations should be changed. In the next section we introduce the new data portrait design, which takes into consideration this change in recommendations, as well as the feedback obtained in the pilot study from the previous section.

Algorithm 5.3 Recommendation of People who Share Intermediary Topics.

```
INPUT: C ← set of candidate users for recommendation
INPUT: u ← target user
INPUT: n ← number of desired recommendations
INPUT: γ ← weight to balance importance between distance and IT similarity
OUTPUT: R ← set of recommended users
  FUNCTION RECOMMEND_USERS(C, u, n, γ)
    FOR ALL c in C DO
          c.distance ← KLD(u, c)
          c.similarity ← JIT(u, c)
    END FOR
      max_distance ← max(C, key='distance').distance
    FOR ALL c in C DO
          c.distance ← c.distance/ max_distance
          c.score ← score(u, c, γ)
    END FOR
      R ← heapq(C, key = score)
    RETURN R.n_largest(n)
  END FUNCTION
```

5.8 A NEW DATA PORTRAIT

In this section we re-design our data portrait paradigm, based on the results from the previous sections.

5.8.1 *Portraying People's Data*

To define our new data portrait design, we have taken into account the user feedback collected during the pilot study. At the core of our design, we maintain our idea of using familiar elements like a wordcloud, augmented with organic, evocative elements. The main change is that now the main element is the wordcloud, instead of tweets as in the previous design. Over the wordcloud there is a histogram that represents time patterns of publishing activity of the portrayed

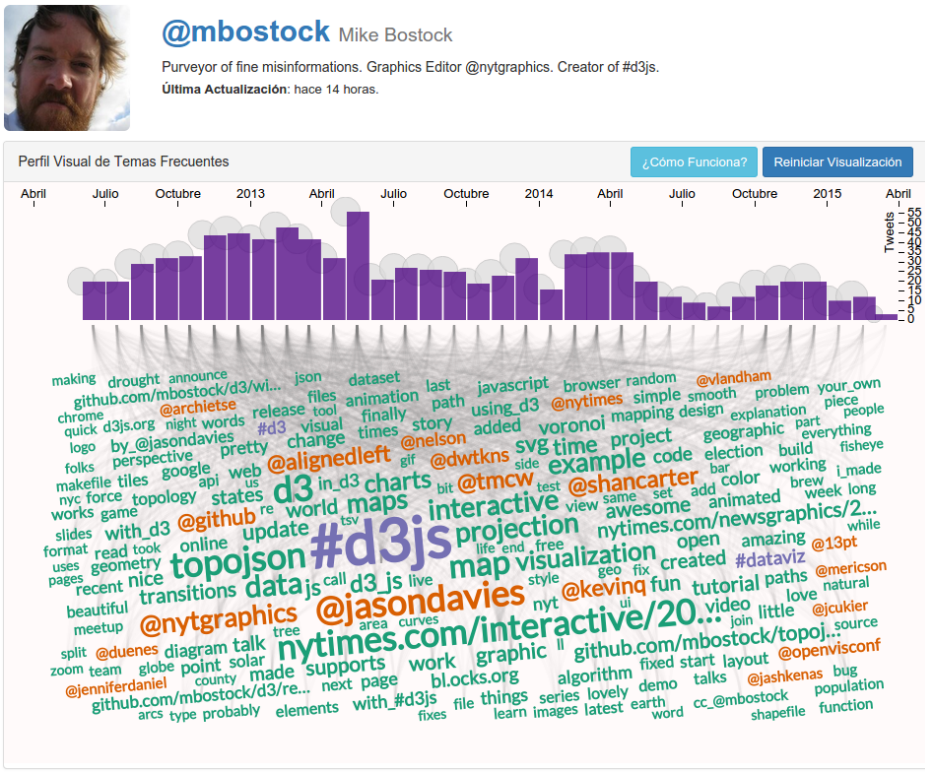


Figure 5.16: New data portrait design. In the image, the portrait of the Twitter account of Mike Bostock, author of d3.js [BOH11].

user, which serves also as a navigation mechanism. Figure 5.16 displays this new design.

Note that the data portrait does not display recommendations. According to the quantitative results in the pilot study, recommendations had a negative effect on self-identification with the portrait. To avoid this effect, in this new design we do not inject recommendations. Instead, recommendations are displayed separately, to make explicit the difference between user interests and recommendations. We describe this representation in the next section, after detailing the rationale behind the portrait design.

Depiction of Users Interests

User interests are estimated in the same way as in the previous portrait design, *i. e.*, by counting frequencies of n-grams (with n up to three). The wordcloud layout is now based on the *Wordle* layout [VWF09], which allows us to have a more tight yet flexible representation of words.[10]

To give a playful appearance to the wordcloud, we applied a rotation to each keyword. Common wordclouds usually follow two patterns of rotation: random rotation; or $90°/-90°$. A random rotation makes hard to read the elements of the wordcloud; the second case gives a sense of structure which is not usually present on the data. Thus, we considered a fixed rotation for all words of $-7°$. This value has been chosen arbitrarily, and we suspect that values between $[-10°, 10°]$ give aesthetically pleasing words.

Likewise our previous design, each word has an invisible box that serves as clickable area, and as indicator when a particular word is highlighted when the box is visible.

Personalization

Although users specified the desire to change the background image of the data portrait, just like they can change the background on the Twitter website, we did not consider background customization because it implies losing control of aesthetics. Instead, we added the user avatar and her/his self description.

Colors and Typography

The color coding of wordcloud elements is based on the type of keyword. We consider three categories: *hashtags* (*#7570b3*), *mentions* (*#d95f02*) and *regular words* (*#1b9e77*). This palette is based on the color-blind friendly *Set2* palette by Harrower and Brewer [HB03]. For typography we use a sans-serif font, as it improves readability [Rel14]. Our current implementation uses the *Lato* font by Google.[11]

10 In particular, we use the implementation by Jason Davies http://www.jasondavies.com/wordcloud/.

11 Available in http://www.google.com/fonts/specimen/Lato.

The Time Dimension

As requested by users, we included time in the data portrait through a histogram of publishing activity. This histogram encodes the number of tweets published or retweeted in a given time window. The number of bins is automatically computed by the *d3.js* library [BOH11]. Each bin of the histogram is accompanied by a circle positioned on its upper-left corner, which serves as an turn-on/off switch of a tweet to be displayed in an overlay window. Although all circles have a similar size, their ratios vary slightly according to the popularity of the most popular tweet of the bin. The usage of this circle allows to select a bin regardless of their size, which is useful specially when some time windows have low activity.

Interactions and Component Linkage

As in our previous design, we link words and bins using *bézier* curves. By contrast, this time the links are always visible, to make the structure behind the data portrait explicit for the user. All links are displayed in a non-highlighted state. To highlight links and change the state of the portrait, the following interactions are available:

- When users click on a specific word, the corresponding bins are highlighted and connected through *bézier* curves (see Figure 5.17 Bottom Left).
- When users click on a specific bin, two things happen:
  - A tweet overlay is displayed with the most popular tweet in it (see Figure 5.17 Top). This tweet is context-dependent: if no word was selected before, it displays the overall most popular tweet; otherwise, it displays the most popular tweet relevant to the corresponding user interest. When a tweet is overlaid, the circle assigned to the current bin is highlighted.
  - The words related to all tweets in the bin are highlighted (see Figure 5.17 Bottom Right).
- When displaying a tweet overlay, if the user clicks the highlighted circle, it is unselected, and the tweet overlay is hidden (see Figure 5.17 Bottom Right).

The state of the portrait can be reset by clicking on a button titled *"Reset Portrait"* (*"Reiniciar Visualización"*). Additionally, we display a button titled *"How*

it Works?" (*"¿Cómo Funciona?"*) that displays a pop-up window with usage instructions.

In this way, we expect users to be constantly exploring different words and bins–moving from related bins (time) to related words (interests), and *vice versa*.

5.8.2 *Displaying Recommendations*

As mentioned earlier in this section, our algorithm generates a list of recommended accounts to follow. In contrast with our pilot study, this time we display *user recommendations* instead of *tweet recommendations*. This means that each recommendation displays a candidate's avatar, the corresponding biography, and a *"Follow"* (*"Seguir"*) button. The full profile in the original platform is linked on the candidate's avatar and username, meaning that the portrayed user can visit his or her profile to get more details about the candidate.

In the application the set of recommendations is displayed below the main data portrait design as a separate unit. However, both are clearly part of the same system.

Using Visualization with Recommendations

We propose that visual depictions have the potential to change how users perceive recommendations. Our rationale has two aspects: first, visualization of social recommendations increases user satisfaction [Gre+10]; second, explaining recommendations is important, as explanations increase user involvement and acceptance [HKR00]. By using visualization techniques to display recommendations, we depict the underlying structure behind them, hence giving an *implicit* explanation. Conversely, when using text only, recommendations have to be explained in natural language, because something like *"Topic 5"* is not meaningful for users. This is an arguably hard problem that is avoided by using a visualization technique.

Circle Packing

We employ *circle packing* [CS03] as a way to display recommendations. We chose to use circles because they maintain aspect ratio (unlike cells in a treemap), which is useful to display avatars at different sizes; and it can be used

to display clusters, as circle packing works with hierarchical data. This is the way it has been used in the past by social query systems like *Hax* by Savage *et al.* [Sav+14]. In our scenario, the clusters can be formed based on the common latent topics between recommended users. In particular, we implemented a simple scheme, where two users were in the same cluster if their most contributing latent topic was the same, although this does not restrict the usage of more complex clustering methods.

The resulting visualization is displayed in Figure 5.18 Top. In contrast with a typical baseline design (see Figure 5.18 Bottom), which displays all recommendations as a list, the circle packing does not display initial recommendations. It shows the corresponding circles and clusters of all users, but no actual recommendation is shown–a message indicating that users can interact with the visualization is displayed instead. When users click on a cluster, the cluster is highlighted and a list on the right of the visualization displays the list of users in that cluster. The format of this list is the same as the one on the baseline interface.

5.9 EVALUATION "INTO THE WILD"

We tested the new design *"in the wild"* [Cra+13], *i. e.*, we deployed an implementation in an uncontrolled setting with end-users. We do so by incorporating both data portrait design and recommender system in the social platform introduced in Chapter 4, *Aurora Twittera* (http://auroratwittera.cl). In *AT*, users could create their "Visual Profiles" (*Perfiles Visuales*) by connecting their Twitter accounts with the site.

5.9.1 *Building the Candidate Set*

Recall from Chapter 4 that *AT* constantly crawls Twitter for tweets about Chilean contingency and news, having in mind geographical diversity. From this always up-to-date dataset we generate, every day, a list of candidate people who have tweeted in the previous 48 hours, and for whom we estimate LDA topics considering the entire corpus of users who published tweets in those 48 hours.

Figure 5.17: State of the data portrait after several interactions. Top: clicking on a histogram bin will display a tweet overlay, with links to all related keywords to that bin. Bottom Left: clicking on a keyword will link all related bins of the histogram. Bottom Right: clicking on the bin circle will deactivate the tweet overlay to ease exploration.

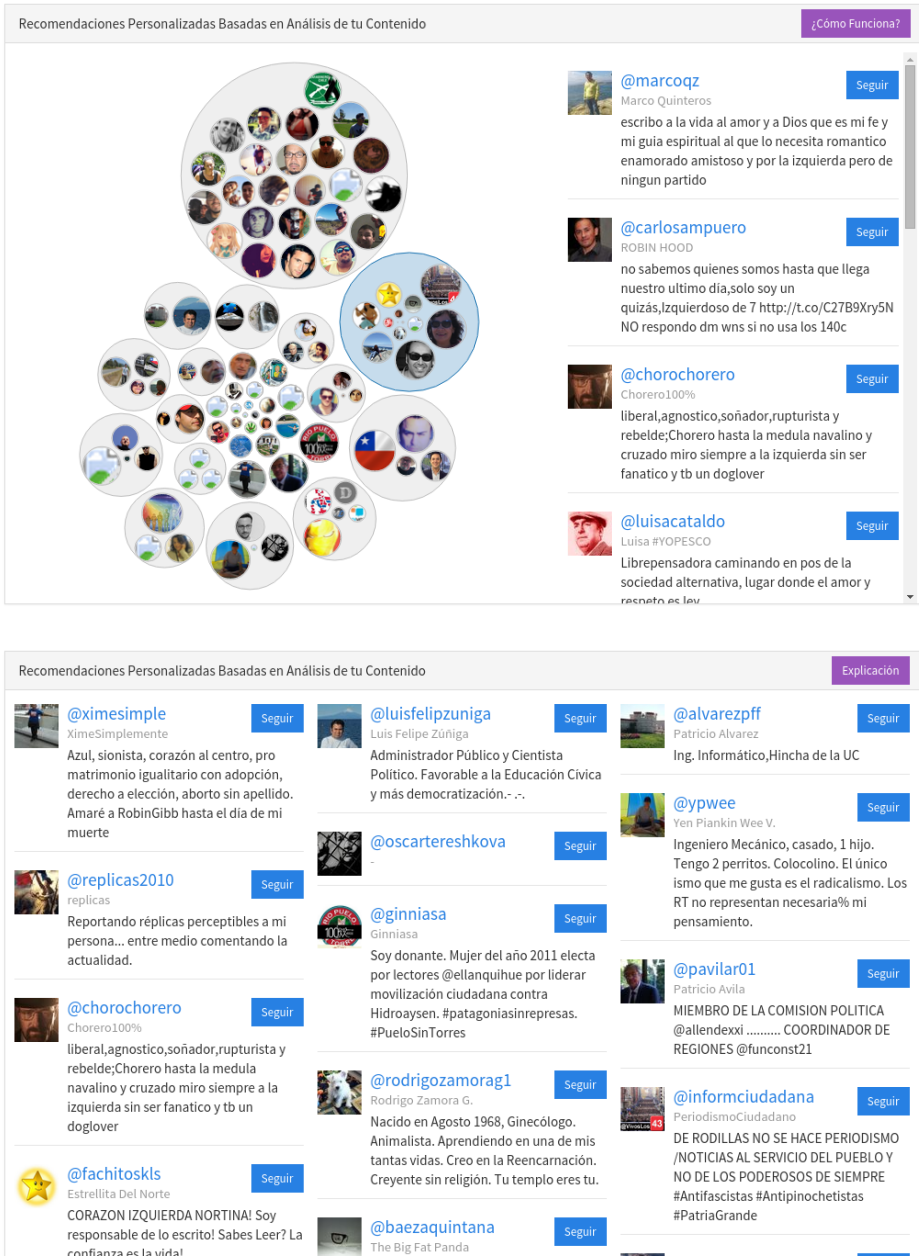


Figure 5.18: Display of recommendations. Top: Circle Packing. Bottom: baseline design.

In our pilot implementation of the recommendation algorithm, user feedback indicated that some recommendations were too old to be relevant. By using this regularly updated list of candidates, we avoided this pitfall and presented fresh recommendations to users everyday.

5.9.2 *Building Portraits and Nudging Users*

When users pressed the *"Create Your Profile"* button, they were redirected to the Twitter website, which asked for login credentials and permission to modify their accounts. We asked for these permissions to be able to have a *"Follow"* button next to each recommendation.[12] Then, a scheduler service processed queued portraits, both those newly created and those queued for update. Its main task was to download tweets using the Twitter API using the credentials given by users. Having downloaded user tweets, we estimated user interests according to our methodology, as well as performing user modeling with LDA.

After their creation, portraits were updated every day, in both user interests and recommendations. Our social bot *@todocl* published tweets mentioning users when their portraits were ready (usually in less than one minute after sign-up), as well as every three days when their portraits were updated. Although updates were daily, notification was limited to every three days per user to avoid spamming.

5.9.3 *Finding Users*

To promote our system we performed the following actions:

- Created several *demo portraits* for people to browse, and publicized them on *@todocl*'s timeline. The demo portraits were about popular user accounts, which sometimes, when being mentioned about the availability of their portraits, retweeted or marked as favorite our announcements.
- Created a campaign on http://ads.twitter.com aimed at Chilean desktop users in Twitter who have been active for at least one month. As re-

12 Unfortunately, the text presented in the Twitter website made several users think that the site wanted to modify their public profiles. This is because permissions are not granular, either it is *read only*, or it is full *read/write*. We believe this had a negative impact in our sign-up rate.

sult, 42,190 promoted tweets were displayed, with an engagement rate (reported by Twitter) of 0.51%.
- Added a "Share my Profile" button to the data portrait. When clicked, the system published a tweet from the portrayed user's account, inviting her/his followers to visit the data portrait.

The system was open to everyone. However, the user interface was available in Spanish only, and recommendations considered only Chilean users as crawled by *@todocl*.

5.9.4 *Interaction Data*

In contrast with our previous experiment (Chapter 4) with interaction data, this time we had rich meta-data associated to each user, with fields such as *full name*, *location*, *self description*, *date of registration*, among others. Following the results of the pilot study, as well as the new design of the recommendation algorithm, we aimed to explore the following questions:

- Does the visualization of recommendations affect how people interact with the recommendations?
- Does the recommendation algorithm affect how people interact with the recommendations?
- Do people who have published political content behave differently (in terms of interaction with the system) than those who do not?
- Do other user characteristics influence behavior, in particular in terms of user engagement?

Experimental Design and Conditions

Our experiment considers a *between-groups* design. The following are the *User Interface* conditions:

- *Baseline*: the baseline recommendation UI (see Figure 5.18 Bottom).
- *Circle Pack*: the visualization of recommendations using circle packing (see Figure 5.18 Top).

The following are the *Recommender System* conditions:

- *KLD*: we computed recommendations using *Topical Closeness* (Kullback-Leibler Symmetric Distance) only.

- *IT*: we computed recommendations using a mixture of *Topical Closeness* (Kullback-Leibler Symmetric Distance) and *Intermediary Topics*, as defined by Algorithm 5.3.

Each condition was randomly assigned to each subject after receiving valid sign-in credentials from Twitter.

Independent Variables

For each subject, we consider the following independent variables:

- *Political Content*: its value is 1 if the list of the top-50 user interests has a non-empty intersection with a list of political keywords (including hashtags); if the intersection is empty, then its value is 0.
- *Account Age*: the number of weeks of the last known activity of the subject since the creation of her/his account.
- *Hub Ratio*: number of friends divided by the number of followers. It is a measure of the tendency of the user to follow others in terms of his/her own popularity.
- *Mention Fraction*: fraction of her/his tweets that mention someone else (not including retweets).
- *RT Fraction*: fraction of her/his tweets that are retweets of others.
- *Tweet Ratio*: number of tweets per day. It is defined as total number of tweets published divided by account age.
- *URL Fraction*: fraction of his/her tweets that contain a URL (without considering retweets).

Dependent Variables

We analyze the following dependent variables:

- *Number of Days*: number of different days the subject visited his/her data portrait.
- *Portrait Events*: number of click interactions with the data portrait.
- *Portrait Sharing*: 1 if the subject shared at least once his/her data portrait on Twitter, 0 if not.
- *Recommendation Events*: number of click interactions with the recommendations (clicking on a profile link, following an account, clicking on circle pack nodes).

- *Dwell Time*: time (in seconds) spent interacting with or browsing the data portrait.

Note that interaction events were normalized by the number of days each participant visited the site.

Procedure

We implemented our data portrait design using the *d3.js* library [BOH11], and the recommendation algorithms using the LDA implementation in the *gensim* library [ŘS10]. These implementations were integrated into the website http://auroratwittera.cl, as well as our social bot *@todocl*, as described in the previous section.

Characteristics of Portrayed Users

Because recommendations considered only Chilean users, and our set of political keywords only considers Chilean political contingency, we discarded users whose self reported Twitter location was not Chilean, or whose IP address was not detected as Chilean by the GeoIP database. We also discarded users whose interaction data was not reliable, for instance, by using ad-blocking software in their browsers (we rely on Twitter Javascript libraries). And, lastly, we discarded users who spent less than 5 seconds on the site.

In total, we have 136 valid portraits, created between February 18th, 2015, and March 17th, 2015. In terms of the recommendation condition, 65 were assigned to *KLD* (Topical Closeness), and 71 to *IT* (Intermediary Topics). In terms of user interface of recommendations, 62 were assigned to the *Baseline*, and 74 to the *Circle Pack* condition. Finally, in terms of political content in their user interests, 73 users had political keywords, and 63 did not. The means of independent variables are: *account age*, 280 weeks; *hub ratio*, 1.29; *mention fraction*, 0.54; *RT fraction*, 0.24; *tweet ratio*, 16.31; and *URL fraction*, 0.17. Figure 5.19 shows the distributions of these independent variables.

5.9.5 *Evaluation Results with Interaction Data*

The 136 portrayed users generated 1801 interaction events with the system, from which we estimated the values of the dependent variables mentioned ear-

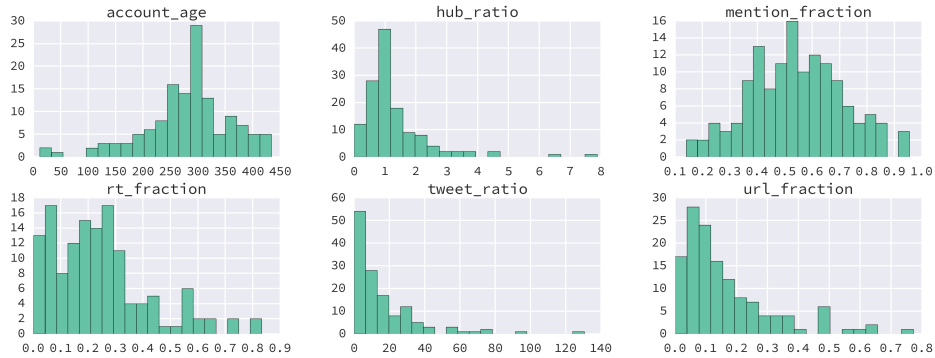


Figure 5.19: Distribution of characteristics (independent variables) of portrayed users.

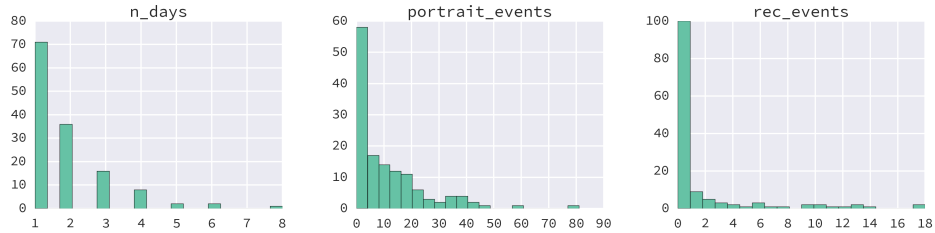


Figure 5.20: Distribution of dependent variables of portrayed users.

lier. Their mean values are: *number of days*: 1.86; *portrait events*: 9.44; *recommendation events*: 1.60. Figure 5.20 shows their distributions. A 52% of participants returned to the site at least for a second day; a 77% of participants shared their portrait in Twitter, either manually by publishing a tweet, or by pressing the *"Share"* button available on the user interface; and 8% of participants accepted at least one recommendation.

Figure 5.21 displays the distribution of *dwell time* without considering the last decile of the distribution. For this variable we discard this decile from analysis because some users left the browser window open.[13]

13 For instance, the maximum dwell time observed was of 49 hours.

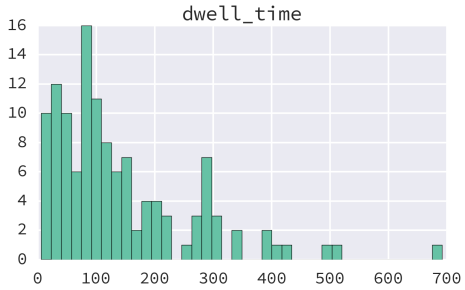


Figure 5.21: Distribution of dependent variable *dwell time* (seconds) of portrayed users. Note that we discard the last decile of the distribution, because some users left their browsers open.

For analysis, we consider the following factorial model:

$$\begin{aligned} Y = {} & C(\text{ui}) \times C(\text{recommendation}) \\ & + C(\text{political\_content}) \times \text{account\_age} \\ & + C(\text{political\_content}) \times \text{tweet\_ratio} \\ & + C(\text{political\_content}) \times \text{hub\_ratio} \\ & + C(\text{political\_content}) \times \text{RT\_fraction} \\ & + C(\text{political\_content}) \times \text{mention\_fraction} \\ & + C(\text{political\_content}) \times \text{URL\_fraction} \end{aligned}$$

Where $C(X)$ creates dummy variables for the corresponding categories of the independent variable X, and $\times$ represent both independent factors and interactions among them. If the interactions were found to be not significant in the model when performing regressions, then we analyzed the same model without interaction terms. When displaying coefficients in Tables we include their value, 95% confidence intervals, and p-value. We identify each result as Ri, to reference it later in discussion. A result can be a mixture of several coefficients in the presence of statistical interactions.

Portrait Events

We performed a Negative Binomial Regression using the model without interactions (scale $= 0.57$; log-likelihood $= -454.21$; deviance $= 111$; $\chi^2 = 72.5$) on

Table 5.6: Negative Binomial Regression Coefficients for *Portrait Events*. *: $p < 0.05$. **: $p < 0.01$.

Variable	β	95% C.I.	p-value
Intercept	1.647	[0.340, 2.954]	0.014*
REC(KLD)	−0.470	[−0.867, −0.073]	0.020*
Pol. Content(1)	−0.486	[−0.886, −0.086]	0.017*
Account Age	0.006	[0.003, 0.008]	0.000***
Hub Ratio	−0.241	[−0.454, −0.028]	0.027*

the variable *portrait events* (interactions were not significant). Table 5.6 displays regression coefficients with a p-value less than 0.05. We observe the following results:

- R1: when users receive recommendations with KLD, portrait events are reduced ($\beta = -0.470$).
- R2: when users have published political content, portrait events are reduced ($\beta = -0.486$).
- R3: users with older accounts tend to interact more with their portraits ($\beta = 0.006$).
- R4: as hub ratio increases, users tend to interact less with their portraits ($\beta = -0.2408$).

Recommendation Events

We performed a Negative Binomial Regression using the model without interactions (scale = 1.08; log-likelihood = −171.74; deviance = 121.89; $\chi^2 = 136$) over the variable *recommendation events* (interactions were not significant). Table 5.7 displays regression coefficients with a p-value less than 0.05. We observe the following results:

- R5: Circle Pack increases interaction with recommendations ($\beta = 2.623$).
- R6: KLD algorithm increases interaction with recommendations ($\beta = 1.357$).
- R7: an increase in tweet ratio decreases interaction with recommendations ($\beta = -0.022$).

Table 5.7: Negative Binomial Regression Coefficients for *Recommendation Events*. *: $p < 0.05$. **: $p < 0.01$. ***: $p < 0.001$.

Variable	β	95% C.I.	p-value
Intercept	−3.3688	[−5.691, −1.046]	0.004**
UI(CP)	2.623	[1.803, 3.444]	0.000***
REC(KLD)	1.357	[0.632, 2.081]	0.000***
Tweet Ratio	−0.22	[−0.044, −0.001]	0.040*
RT Fraction	3.241	[1.203, 5.280]	0.002**

Table 5.8: Logistic Regression Coefficients for *Has Accepted Recommendations*. *: $p < 0.05$. **: $p < 0.01$. ***: $p < 0.001$.

Variable	β	95% C.I.	p-value
REC(KLD)	2.835	[0.730, 4.939]	0.008**
Pol. Content(1)	2.163	[0.306, 4.019]	0.022*
Mention Fraction	−6.844	[−12.147, −1.540]	0.011*

- R8: an increase in RT fraction increases interaction with recommendations ($\beta = 3.241$).

Recall that 8% of participants accepted at least one recommendation. We performed a logistic regression over this outcome (Pseudo $R^2 = 0.30$, log-likelihood $= -26.70$, $p = 0.006$; the model with interactions was not significant). Table 5.8 displays regression coefficients with a p-value less than 0.05. We observe the following results:

- R9: KLD algorithm increases the likelihood of recommendation acceptance ($\beta = 2.835$).
- R10: users with political content are more likely to accept recommendations ($\beta = 2.163$).
- R11: an increase in mention fraction decreases likelihood of accepting recommendations ($\beta = -6.844$).

Table 5.9: Negative Binomial Regression Coefficients for *Number of Days*. *: $p < 0.05$. **: $p < 0.01$.

Variable	β	95% C.I.	p-value
Pol. Content(1)	2.287	[0.765, 3.809]	0.003**
Hub Ratio	0.150	[0.000, 0.299]	0.049*
Hub Ratio and Pol. Content(1)	−0.299	[−0.574, −0.024]	0.033*
Mention Fraction	1.031	[0.0402.023]	0.041*
Mention Fraction and Pol. Content(1)	−1.859	[−3.206, −0.512]	0.007**

Number of Days

We performed a Negative Binomial Regression over the variable *number of days* (scale = 0.16; log-likelihood = −278.50; deviance = 14.852; $\chi^2 = 18.4$; there were significant interactions). Table 5.9 displays regression coefficients with a p-value less than 0.05. We observe the following results:

- R12: users with political content are more likely to revisit the site ($\beta = 2.287$), but this likelihood is reduced with increments in users' hub ratio (interaction with $\beta = -0.299$) and mention fraction (interaction with $\beta = -1.859$).
- R13: an increase in hub ratio increases likelihood to revisit the site ($\beta = 0.150$).
- R14: an increase in mention fraction increases likelihood to revisit the site ($\beta = 1.031$).

Portrait Sharing

We performed a logistic regression over the variable *portrait sharing* (Pseudo $R^2 = 0.05$, log-likelihood = −69.30, $p = 0.59$; the model with interactions did not have significance). Even though the fit is poor, the coefficient for URL fraction ($\beta = 3.546$, 95% C.I. [0.559, 6.534], $p = 0.020$) indicates that, as users have a greater URL fraction, the likelihood to share the portrait increases (R15).

Table 5.10: Gamma Regression Coefficients for *Dwell Time*. *: $p < 0.05$.

Variable	β	95% C.I.	p-value
Intercept	0.012	$[-0.002, 0.0026]$	0.095
UI(CP)	0.0063	$[0.000, 0.012]$	0.046*
UI(CP) and Pol. Content(1)	-0.0081	$[-0.015, -0.001]$	0.023*
UI(CP), Pol. Content(1) and REC(KLD)	0.011	$[0.001, 0.021]$	0.032*
Account Age	-2.53×10^{-5}	$[-5.05 \times 10^{-5}, -1.47 \times 10^{-7}]$	0.049*

Dwell Time

We performed a Gamma regression over the variable *dwell time* (log-likelihood $= -709.93$, deviance $= 78.96$, $\chi^2 = 63.3$).[14] Recall that we discard the top decile of the distribution, thus $N = 102$. Table 5.10 displays regression coefficients with a p-value less than 0.1. We observe the following results:

- R16: Circle Pack decreases dwell time ($\beta = 0.0063$). However, the effect is simple and counter-effected when users are politically involved ($\beta = -0.0081$).
- R17: the mixture of Circle Pack, KLD recommendations and Presence of Political Content, decreases dwell time ($\beta = 0.011$).
- R18: an increase in account age increases dwell time ($\beta = -2.53 \times 10^{-5}$).

5.9.6 *Overview of Findings and Discussion*

Here we analyze the obtained results in terms of the questions asked before the evaluation.

What affects interaction with the portrait?

The following factors influence interaction with our portrait design:

- *Recommendation Type*: portrait events were reduced when users received KLD (baseline) recommendations (R1). However, this does not mean less

14 Recall that parameter interpretation in Gamma is different from the other models used, because we use an *inverse power* link function. Negative coefficients imply positive contribution.

engagement with the application, as those users were more likely to generate recommendation events (R6).

- *Presence of Political Content*: users with political content in their portraits generated less portrait events (R2). Yet, those users were more likely to revisit the site (R12), and thus, we observe that they are engaged differently with the site than those users who do not have political content.
- *Network Activity*: a higher hub ratio implied less portrait interaction (R4), which corresponds with the interaction of hub ratio and political content that reduced likelihood to revisit the site (R12). This influence by hub ratio makes sense, given that users with hub ratio greater than one are those who have more friends than followers, which can be interpreted as a passive way to use Twitter–they are users who could be more interested in reading other's tweets than to build a network of followers.
- *Account Age*: users with older accounts were more likely to generate portrait events (R3). This is expected, as older users have more content to be portrayed, and it is reasonable to assume that they would be interested in exploring their accounts.

What affects interaction with recommendations?

The following factors influence interaction with the recommendation part of the system:

- *Recommendation Type*: we found that recommendation events increased when users received KLD recommendations (R6). This is expected, as our initial assumption is that users prefer this kind of recommendation over more politically diverse ones. In fact, in concordance with homophily, KLD increases likelihood of acceptance (R9).
- *Visualization*: when recommendations are displayed with Circle Pack, recommendation events also increase (R5). This means that, regardless of recommendation algorithm, users are more likely to be exposed to the recommendations when they are visualized.
- *Publishing Behavior*: behavioral factors that influence recommendation exploration are tweet ratio, which decreases events (R7), and RT fraction, which increases events (R8). Recall that tweet ratio is the average number of tweets per day, meaning that someone who publishes lots of content is

either generating it, or has already a network to gather content from. Arguably, this high level of activity is more about generating content than otherwise. Conversely, in the case of RT fraction, a higher fraction implies more content promoted from other sources, which means that the portrayed user values information from others. Our results indicate that such users generate more recommendation events, perhaps to find more sources to retweet from. Similarly, when mention fraction increases, likelihood of recommendation acceptance decreases (R11), possibly because those users have already built their networks, and as such are focused on interacting with it rather than building it.

- *Presence of Political Content*: recommendation acceptance likelihood increases when users have political content in their portraits (R10). This extends our finding from the pilot study, where users who had tweeted about abortion gave better ratings to recommendations.

Which users are engaged?

We observed that 52% of participants returned to the site at least a second time in a different day; and that 77% of participants shared their portrait in Twitter. Based on results, we identify the following factors that influence user engagement:

- *Presence of Political Content and Informational Behavior*: politically vocal users had a greater tendency to return to the site, although hub ratio and mention fraction decrease this tendency (R12). Both characteristics are related to the number of people a participant interacts with, either by following more users than being followed, or by mentioning/replying more, such that mentions have an important fraction of the user's timeline.
  In terms of hub ratio, we can analyze this result in the context of *informer users* [NBL10], *i. e.*, those users whose main role is to spread information. *Informers* are focused on attracting followers, and thus their hub ratio is expected to be low. In terms of higher mention fraction, given that a considerable amount of their content is interaction with others, perhaps their discussion networks are already built and they visit the site once to confirm the image they project [Gof59]. In both cases, when users are politically involved, once they have seen their portraits there is no need to

return, as political stances are expected to have, if any, a low variance in time. The opposite happens when users are not politically involved (R13 and R14). Those users are informing about non political issues, which arguably present more variability in time than political contingency.
In our pilot study, some users expressed that data portraits are useful to confirm if curated content is projecting the desired image; we devise this result as an expression of that qualitative feedback.

- *Publishing Behavior*: users who have higher URL fraction are more likely to share their portraits (R15), providing an additional link between engagement and *informers*.
- *Account Age*: users with older accounts were more likely to exhibit greater dwell time (R18). This is in concordance to their greater likelihood to interact with the portrait (R3).

Because of its complexity, we discuss separately the interaction of *Visualization, Presence of Political Content and Recommendation Type* found in the analysis of dwell time.

On one hand, Circle Pack, regardless of the recommendation algorithm, reduces dwell time (R16). Moreover, when using KLD jointly with Circle Pack and users are politically involved, dwell time is also decreasing (R17). Given that users performed more recommendation events with Circle Pack (R5) and KLD (R6), this means that they perform their exploration of recommendations faster when using the Circle Pack. These users exhibit a *focused* recommendation exploration.

On the other hand, when Circle Pack is presented to politically involved users, the effect is opposite and dwell time increases (R16). Given (R17), this means that Circle Pack jointly with Intermediary Topics increase dwell time of politically involved users. Considering the following previous results:

- Politically involved users are more likely to accept recommendations (R10).
- Visualization is likely to increase recommendation events (R5).
- Intermediary Topics users are less likely to interact with recommendations than their KLD counterparts (R6).

Then, we can conclude that politically involved users who were exposed to diverse recommendations using visualization exhibit a *reflective* recommendation exploration.

5.10 DISCUSSION

In this section we discuss the overall results from this chapter. We started with the premise of connecting people with opposing views, then performed a first design proposal of both an algorithm to recommend such people, and a novel user interface to display those recommendations.

Initial Methodology and Case Study: Abortion in Chile

To put our proposal into context, as well as to validate our motivation, we performed a case study on a subset of the Chilean virtual population in Twitter. Specifically, we targeted the sensitive issue of abortion. The contingency at the time of the study generated debate on the issue, and discussion around abortion exhibited *homophilic behavior*. This scenario was ripe to test our first design, and we performed a pilot user study.

Pilot Study

Results of this study shown that our initial implementation of recommender system did not work as users expected. Since our algorithm was based on a search engine using the vector space model [BR11a], sometimes recommendations were relevant in terms of keyword matching, but not on the meaning of those keywords. However, we still obtained deep insights in terms of user behavior. Particularly, the key result we found is that recommendation evaluation was influenced by the self-reported answer to *"Have you tweeted about abortion before?"*. This content-based difference indicated that political openness might be a feature that could help us to differentiate users according to their behavior.

To surpass the limitation imposed by the vector space model, we improved our algorithm by using Latent Dirichlet Allocation [BNJ03], and then we developed the concept of *Intermediary Topics*. We extended the case study to perform an off-line evaluation of intermediary topics, and found that they exist, that they can be measured, and that they can be used to provide recommendations of people of opposing views.

New Data Portrait Design

Considering feedback from the pilot study, we introduced a new data portrait design where we separated the depiction of recommendations from the data portrait itself. This allowed us to test a visual strategy to depict recommendations without changing the core data portrait design, as well as finding how recommendations influenced interaction with the data portrait.

Results of Deployment "In the Wild"

We evaluated our designs considering interaction data of Chilean users who registered on our *Aurora Twittera* platform, obtaining interesting results about users and their behavioral signals.

In one hand, behavioral and content differences influence how users perceive a *Casual Information Visualization* system [PSM07]. We found that being politically open, interacting with others and being an *informer* are important features that influence interaction. This is important because, when designing InfoVis systems for specific tasks, user characteristics can be assumed by visualization designers, but in open systems like ours we cannot predict who will use the system nor their expertise level. Knowing which behavioral and content signals influence their usage of the system will help to design, and even customize, new interfaces.

In the other hand, there are complex interactions within the conditions and variables studied. We can say that the usage of visualization encouraged *exploration* of recommendations, regardless of the algorithm used to generate them. In contrast, if we look at recommendation *acceptance* from a general perspective, people still behaved in an homophilic way. However, considering exploratory behavior according to *dwell time* leads to deeper insights. We observed that politically involved users who received diverse recommendations depicted visually, *i. e.*, those who were affected by our proposed conditions, performed a *reflective exploration*. Because politically involved users were more prone to accept recommendations, and our proposed conditions made those users spend more time with the system, then we can conclude that those users shown a conscious decision-making process, *i. e.*, their behavior avoided the cognitive heuristics that lead to biased behavior.

5.10.1 *Implications*

We contextualize the implications in three areas, each one with a key question: *individual differences* (*who?*), *recommendations* (*when?*), and *data portraits* (*how?*).

Modeling and Detecting Individual Differences

Our results indicate that individual differences matter, both at the perception level (*i. e.*, recommendation receptivity) as found in the pilot study, as in the application level (*i. e.*, interaction behavior with portrait and recommendations), as found in the evaluation of interaction data of end-users. We focused on behavioral signals that could be extracted from user profiles, namely, publishing behavior (*tweet ratio*), experience (*account age*), connectivity (*hub ratio* and *mention fraction*), and *informational behavior* (*RT and URL fraction* of their tweets). By modeling those characteristics with a statistical model, including an interaction of those signals with the detection of political content in user interests, we were able to differentiate behavior between politically vocal and politically silent people. This was an important distinction, as we found that these signals had significant effects on user engagement with the system, including interactions with the presence of political content in a user's portrait.

Thus, even classifications of users who are not about politics, like *informers* and *meformers* [NBL10], need to account for politically vocal users, because *political informers* behave differently than *non-political informers*. This makes sense, as, in line with our motivation, arguably only political informers are affected by selective exposure, in the sense that they look for political content, while non-politically involved people discards political content because of lack of interest instead of selective exposure. Not all users are interested in politics, therefore, not all users are interested in, nor need, political diversity on their timelines.

Recommending Diverse Content

Our premise was that diversity-aware recommendations had the potential to activate selective exposure in users, who would, in turn, discard recommendations. To avoid selective exposure, we proposed to rely on *intermediary topics*. Then,

we wanted to encourage a positive reception by users, and we proposed to do so by using an aesthetically attractive design based on circle packing [CS03], which had the property of displaying part of the underlying structure in recommendations. This indirectly tackled the hard problem of explaining recommendations [HKR00]; since our recommendations are based on Latent Dirichlet Allocation [BNJ03], explanation is difficult to achieve, as latent topics are mathematical and do not always make sense in common language.

Although users did not interact with recommendations as much as they did with the data portrait, the captured events were enough for our model to explain part of user behavior from a quantitative point of view. Users behave in homophilic way, by interacting more with recommendations when they were generated by the baseline algorithm. In this aspect, we observed that our visualization proposal was effective: when visualizing recommendations instead of using the baseline text interface, users' exploration of recommendation was equivalent, or even greater, than when recommendations were non-diverse. This implies that *visualization is effective to encourage exploration of recommendations.* This result is similar to previous work by Faridani *et al.* [Far+10], where visualization improved the behavior of users discussing sensitive issues, as they were more respectful with others with opposing views, although it had no impact on selective exposure. In that aspect, we did not found a main effect of visualization in reducing homophilic behavior of recommendation acceptance. In fact, homophilic behavior was confirmed, because the non-diverse recommendation algorithm increased likelihood of acceptance. This result is not surprising–this is why we acknowledge that no two users are equal, and why we have discussed and analyzed individual differences.

To understand *when* our proposed approach of visualization and intermediary topics work, we characterized exploratory behavior in terms of *dwell time* with our application. A differentiating attribute in this aspect was whether users performed a *focused* or a *reflective* exploration of recommendations. As result, we found that, when users are politically involved, *visualization and intermediary topics are effective to increase thoughtful decision-making on recommendations.*

Data Portraits as a Tool for Engagement

We acknowledge that non-expert users might exhibit resistance when faced to disruptive changes in the interfaces they are used to, such as the text based interfaces present today in social networks, micro-blogging platforms, and media outlets. Those interfaces have a cold business-like feeling and do not represent the lively social context they are supposed to represent [Don14]. To be able to change this feeling into a more evocative one, but without introducing disruptive changes, we added a new stimuli to a familiar element. We augmented in a substantial way the common interaction with wordclouds, while providing a friendly and evocative appearance based on organic design.

We theorize that the main cause of the positive engagement found is our data portrait design, because we have followed the feedback obtained in the pilot study. The results discussed in terms of interaction with the portrait and recommendation, plus the positive engagement partially explained by our model, implies that *current systems should introduce optional casual information visualization user interfaces to explore content generated by users*. This would have several benefits for users, as they will be able to unleash their social potential by being able to choose the user interface that suits best for their exploratory styles.

5.10.2 *Summary, Limitations and Future Work*

As mentioned by Kevin Lynch in his book "What Time is This Place?", *"the best environment for human growth is one in which there are both new stimuli and familiar reassurances, the chance to explore and the ability to return"* [Lyn72]. In this chapter we have tried to define one of such environments to explore user generated content in micro-blogging platforms, where we considered human growth as the quality of being able to connect with others of opposing views. In particular, our proposed environment was a *Casual Information Visualization* [PSM07] system that implemented the *Data Portraits* paradigm [Don+10]. Through a process of iterative design, we opened a path where the design of user interfaces jointly with algorithms that exploit user biases have the potential to change the way people interacts with exploratory systems. Because we presented an exploratory system, our evaluation was not a task-based one focused

on algorithm/visualization efficiency. Instead, we performed an "in the wild" evaluation [Cra+13], where we focused on individual differences and user interaction with the interface, as well as user engagement metrics [LOY14]. Such focus allowed us to obtain deep insights on user behavior and exploratory styles.

After discussing our results, we outlined the implications of our work. In particular, by focusing on individual differences we have understood part of *who* should be targeted with proposals like ours. By focusing on the recommendations aspect, we have understood *when* should paradigms like ours be used. And finally, by focusing on the information visualization aspect of this project, we have understood part of *how* our proposal can lead to good results. Yet, not all is done–there are still many questions, in particular in terms of the *why* behind our results. These aspects have been covered before in data portraits scenarios by Viégas, Golder, and Donath [VGD06] with qualitative studies, although their setting was different and non conflictive. This is what future work awaits.

Limitations

There are three main limitations of this work. First, even though the number of participants was sufficient for our model to reliably perform regressions using generalized linear models, a greater number of participants would have allowed us to present more evidence to support our claims. We believe we could have obtained more sign-ups if Twitter permissions were more granular, as in their current form they exceed what our application really needs, and this scared users. Second, our recommendation algorithm was only content-based, without considering network features as input, something that is expected by users [Che+09]. Network features also aid when defining explanations (*e. g.*, by making common contacts explicit). Finally, our design focused on desktop environments, but now mobile platforms are more common, specially in non-expert users, and thus, coverage of mobile scenarios is needed.

Future Work

In addition to addressing our limitations, we plan to perform a qualitative evaluation. In particular, we consider a framework that will help to understand the *why* on the results like *PLayful EXperiences* by Lucero *et al.* [Luc+13]. By conducting such evaluation, we will be able to perform a similar analysis to the one

by Viégas, Golder, and Donath [VGD06]. We expect to, in addition to analyze insights and usage of the portrait, find how users perceive privacy with respect to their content–a data portrait surfaces the patterns behind user generated content, and users are not always conscious about the information they publish on the Web.

6

CONCLUSIONS

In this dissertation we studied the effects of bias in behavior of Web users by following a transversal approach based on case studies. In this chapter we discuss the results and implications of the dissertation as a whole, as well as future work to be pursued.

6.1 SUMMARY

We performed three case studies in two different platforms. We started with Wikipedia, focusing on community maintained content (Chapter 3), and then we moved to Twitter, a micro-blogging platform where we studied behavior in terms of the user generated content (Chapters 4 and 5). The case studies performed are:

- *Gender Bias on Wikipedia*, where we focused on understanding, quantifying, and contextualizing the bias introduced by the community in the depiction of women.
- *Political Centralization on Twitter*, where we focused on how a systemic bias from the physical world affects the entire Web content lifecycle.
- *Political Homophily on Twitter*, where we focused on how a cognitive bias affects how users communicate with others.

Next, we summarize the results of each case study.

6.1.1 *Gender Bias on Wikipedia*

In Chapter 3 we studied *gender bias* in user generated content on the open encyclopedia Wikipedia. We found that there are biases in how women are characterized in biographies in comparison to men, as well in the network structure of links between biographies. These biases were quantified and their consequences on content were identified. For instance, women are harder to find than men because the network structure is biased in terms of centrality in the network of biographies. We contextualized the differences in characterization in terms of social theory from a feminist point of view. This allowed us to see that not all differences correspond to biases of the Wikipedia community, as some differences were reflections of bias in society. Based on our findings, we suggested several ways to improve Wikipedia's current guidelines. On one hand, we suggested to implement new guidelines of bias in language, as the current neutral point of view guideline does not inhibit such bias. On the other hand, we inquired Wikipedia to perform affirmative actions to relax their guidelines of notability for women, given how women have been excluded from history and are harder to find.

6.1.2 *Political Centralization on Twitter*

In Chapter 4 we studied *political centralization* and its effect on Twitter in the context of a centralized country. We analyzed the entire Web content life cycle, *i. e.*, from content-generation by users to content-consumption, including content classification by machine learning algorithms. We found that centralization affects all aspects of the life cycle, including user behavior.

To reduce bias effects and give users the chance to explore an unbiased timeline, we built an information filtering algorithm and an user interface based on visualization techniques. We evaluated the algorithm and the user interface with users in qualitative and quantitative terms, and found important differences in user perception induced by the systemic bias.

Then, guided by concepts of identification drawn from ethnographic literature, we implemented a design based on treemaps. We found that its mixture with our information filtering algorithm enabled users to explore unbiased time-

lines in an "in the wild" implementation. Furthermore, we found that users from central locations behave differently when interacting with the system than those users from peripheral locations.

6.1.3 *Political Homophily on Twitter*

In Chapter 5 we studied *homophily* and its effect on Twitter, in particular in the context of the politically-involved population of a country. In a case study which analyzed the discussion about abortion, we found that people interact with others in a homophilic way in terms of positions on abortion. We proposed an initial algorithm to recommend people of opposing views, and an initial design for a data portrait of users, which would serve as a non-political context to present recommendations. Those recommendations of people with opposing views, were generated from an analysis of shared interests that were non-conflictive. In this way, we used homophily in one aspect to encourage reduction of homophily in another.

We ran a pilot study to evaluate our initial proposals. We found that there were individual differences in perception of user recommendations in terms of political-involvement of users. User feedbacks allowed us to improve the design of the system as a whole, which we tested "in the wild". Results of this deployment were mixed. On one hand, we found that homophilic behavior is strong and users, in general, accepted more recommendations when they were fully homophilic. On the other hand, we found positive user engagement with the application and we confirmed our proposal that visualization encourages exploration of recommendations. Furthermore, we found that when users are politically involved, the mixture of visualization and our intermediary topics not only encourages them to explore. It also enables a *reflective exploration process*, which we link with a conscious decision-making in terms of recommendations, *i. e.*, an unbiased, rational behavior.

6.2 CONTRIBUTIONS AND IMPLICATIONS

In this section we analyze the contributions and implications of the main results derived from the transversal analysis of our case studies.

6.2.1 *Understanding Biases in Content and Behavior*

Web Mining Tools are Effective when Measuring Bias in Content and Behavior

We started inquiring if biases from the physical world are reflected on the Web. Through the three case studies we confirmed the presence and effects of the specific biases under consideration. In Chapters 3, 4 and 5 we observed, analyzed and quantified behavior by employing techniques drawn from Computational Linguistics, Information Retrieval, Topic Modeling, Machine Learning, and Network Analysis. We used those techniques to understand user generated content and usage of Web platforms, as well as our own deployed systems "in the wild", in the context of the Web Mining process described in Chapter 2.

Researchers wanting to understand biased behavior in Web activities will be able to do so by employing known techniques related to Web Mining. These techniques will be useful in any part, or even in the whole, of the Web content life cycle.

Social Sciences Frame and Guide the Analysis

Although this dissertation is written in a Computer Science context, Web Mining tools only provide a quantitative framework; the question of *what to measure* cannot be answered solely by Computer Science. In this aspect, in all chapters we have guided the analysis and proposed designs by relevant Social Science theory: *feminism* in Chapter 3, *ethnography* in Chapter 4, and *homophily* and *presentation of self* in Chapter 5. This is why we contextualized this dissertation in the field of Computational Social Science.

By having a Social Science framework to base the analysis, researchers can evaluate which factors are important when defining the rationale that will guide algorithmic output and user interface design of exploratory and Casual InfoVis systems, as we have done in Chapters 4 and 5.

6.2.2 *Encouraging Changes in Behavior*

Unbiased Algorithms are Necessary but not Sufficient

Whether users value or see diversity is something that depends on the specific bias being analyzed. In Chapter 4 we focused on a systemic bias which affected user perception and behavior. Arguably, this is something that might be out of control to the user, or the user just do not care about it because of conformism with the system. In Chapter 5 we focused on a cognitive bias, in which users chose to have biased behavior because it was beneficial for them (as they avoided cognitive dissonance). Then, although an algorithm can theoretically diminish a bias by generating diverse output (*e. g.*, recommendations or timelines), it is naive to think that algorithmic design by itself will make users explore content in a less biased manner. Just delivering different information will not encourage a change in their behavioral choices.

Researchers wanting to encourage change in behavior in biased contexts must think how to complement the necessary algorithms that deliver unbiased content, for instance, by designing new user interfaces.

However, note that algorithms and quantification can be sufficient if the purpose is just quantification of bias and not encouragement behavioral change. For instance, although we did not focus on changing biased behavior of Wikipedia editors in Chapter 3, our results still made impact in the community, who adopted our resulting guidelines for writing about women.[1]

Information Visualization Encourages Exploration of Diverse Content

We have found that the usage of information visualization techniques encourages exploration of diverse content. Note that we refer to specific visualization techniques, in particular, we focused on hierarchical techniques like *treemaps* and *circle packing*, instead of trying several different ones. To find if visualization effectively encouraged exploratory behavior, we measured differences with respect to specific baselines pertinent to current mainstream interfaces. Our rationale is that we chose which techniques to use based on the literature as well

1 http://en.wikipedia.org/wiki/User:GGTF/Writing_about_women

as the entire context surrounding the case study. Future work may wish to revisit this idea and evaluate more visualization techniques in these contexts.

In terms of the purpose of our work, as noted in the introduction of this dissertation, our designs aimed to have three traits as defined by Donath [Don14]: *to be innovative*, *to be legible*, and *to be socially beneficial*. We tried to be innovative by approaching every bias from an unexplored angle, guided by the insertection of Web Mining, Social Sciences and Information Visualization; we tried to be legible by designing visualizations having end-users and their social context in mind when defining user interface rationale; and we tried to be socially beneficial by focusing on the benefits of having a less biased behavior.

Whether we succeeded in carrying those traits is something time will tell. However, as discussed, the results from Chapters 4 and 5 confirm that usage of Information Visualization, in particular in the context of Casual InfoVis systems, produced the wanted effect of encouraging exploration. Therefore, a visualization designer who wants to encourage exploration of diverse/unbiased information should consider a design process guided by those traits.

One Size does not Fit All

Culture, Geography, Social context and individual differences matter when studying user behavior. No two users are equal, and, as such, these factors must be considered when designing exploratory systems in biased scenarios. In this aspect, Bardzell [Bar10] provides a framework for Human-Computer Interaction against universal design, because universal points of view in artifact design become normative and totalitarian. For instance, in Chapter 5 we found that different archetypes of users in terms of informational behavior have opposite behaviors in biased scenarios, and thus, a universal approach would not have encouraged unbiased behavior at all.

Researchers must consider that, while systemic and cognitive biases affect people in general, everyone is affected in different ways.

User Engagement Allows to Measure Differences in Behavior "In the Wild"

Casual InfoVis and exploratory systems are hard to evaluate, as they are not task-based, and thus, metrics like accuracy, efficiency and task-time are not available nor meaningful. In Chapters 4 and 5, user engagement metrics like

return probability and dwell time were used as measured in our scenarios. In addition to behavioral and informational signals, they allowed to measure differences in behavior, enabling us to effectively evaluate if a Casual InfoVis design was effectively used as intended, and if it had encouraged the intended changes in behavior.

This is an important contribution, because it allows to perform a quantitative evaluation of these kind of systems. For instance, data portraits (see Chapter 5) have been evaluated mostly in qualitative manners, which gives deep understanding of usage but does not allow to study qualities like encouragement of behavioral change nor exploration by users according to individual differences. By using user engagement metrics, researchers can perform "in the wild" studies to obtain quantitative results, which can be later explained with qualitative studies.

6.3 FUTURE WORK

In this section we outline several lines of work to be pursued after this dissertation.

6.3.1 *Replication Instead of Generalization*

We believe our results should not be generalized to the entire Web, not only because we focused on specific case studies which might not be representative, but because we started from a premise of cultural and social differences. Following that premise, instead of generalization we will seek *replication* of our studies in other cultural and social contexts.

6.3.2 *A Framework for Evocative Visualization in Biased Scenarios*

We defined design guidelines in our case studies based on Social Sciences contexts, as well as user feedback. However, we did not define a conceptual framework for evocative visualizations that encourage exploration of diverse content in biased scenarios. Such framework will be helpful for visualization practitioners and designers when creating new visualizations in social Web platforms. To

implement this, it would be pertinent to try different visualization techniques from the ones we have used.

6.3.3 *Exploratory and Interactive Contexts for Longitudinal Studies*

The presented systems, although evaluated "in the wild", are still proof of concept implementations that do not allow users to embrace our proposed designs as replacements of the original platforms we targeted. For instance, the *treemap* visualization from Chapter 4 does not display conversations around the filtered tweets, and, although a tweet can be displayed in its native format, subsequent actions performed on it are not reflected on the user interface. Hence, by having a more integrated exploratory experience with extended interaction mechanisms, we will be able to run longitudinal studies that will allow us to perform deeper analysis in our context.

6.3.4 *Mobile Contexts*

Finally, we will consider mobile platforms as target for our next studies. Undoubtedly, today the mobile context is crucial when studying behavior of end-users, in particular non-experts as targeted by this dissertation. The mobile context delivers new opportunities, as usage contexts are more varied than desktop ones, as well as new challenges, because design guidelines for desktop environments do not apply.

6.4 FINAL WORDS

In this dissertation there was an implicit common concept through all the case studies: *the other*. In the first case study, the other was women. In the second, it was those who do not live in the capital of a centralized country. In the third, it was those who think differently. The biases in our minds and in our systems are making us diminish, ignore or avoid them. By doing so, we are missing a world full of potentially enriching views. Sometimes, such thinking is not conscious. As we have found in this dissertation, it is possible to design algorithms and user interfaces that allow users to make conscious choices. To do so is necessary to

become empathic in design. In Computer Science it is hard to be empathic as a designer (either of algorithms, user interfaces or systems), because diversity is not a common trait found neither in the academy nor in the industry. There are efforts to change this, but as long as we, as computer scientists, do not embrace diversity in *what* we make, in addition to *where* we make it, then the Web will be of only one color.

A globalized Web has plenty of opportunities for human growth; an uniformized Web has not. We hope to have planted a seed toward a Web full of color.

BIBLIOGRAPHY

[Abb+09] Zeinab Abbassi *et al.* "Getting recommender systems to think outside the box". In: *Proceedings of the third ACM conference on Recommender systems*. ACM. 2009, pp. 285–288.

[Abe+12] Fabian Abel *et al.* "Semantics+ Filtering+ Search= Twitcident. Exploring information in social web streams". In: *Proceedings of the 23rd ACM conference on Hypertext and Social Media*. ACM. 2012, pp. 285–294.

[ABP14] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. "powerlaw: a Python package for analysis of heavy-tailed distributions". In: *PloS ONE* 9.1 (2014), e85777.

[Abr75] Jane Abray. "Feminism in the French Revolution". In: *The American Historical Review* (1975), pp. 43–62.

[AD09] Yannick Assogba and Judith Donath. "Mycrocosm: visual microblogging". In: *42nd Hawaii International Conference on System Sciences*. IEEE. 2009, pp. 1–10.

[AG05] Lada A Adamic and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog". In: *Proceedings of the 3rd international workshop on Link discovery*. ACM. 2005, pp. 36–43.

[AHS13] Amr Ahmed, Liangjie Hong, and Alexander J Smola. "Hierarchical geographical modeling of user locations from social media posts". In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 25–36.

[Aie+12] Luca Maria Aiello *et al.* "People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks". In: *Links* 697.483,151 (2012), pp. 1–566.

[ALR12] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors." In: *International Conference on Weblogs and Social Media* 270 (2012).

[AlS+09] Loulwah AlSumait *et al.* "Topic significance ranking of LDA generative models". In: *Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 67–82.

[AMC07] Rodrigo Almeida, Barzan Mozafari, and Junghoo Cho. "On the Evolution of Wikipedia." In: *International Conference on Weblogs and Social Media*. 2007.

[An+14] Jisun An *et al.* "Sharing political news: the balancing act of intimacy and socialization in selective exposure". In: *EPJ Data Science* 3.1 (2014), pp. 1–21.

[Ant+11] Judd Antin *et al.* "Gender differences in Wikipedia editing". In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM. 2011, pp. 11–14.

[APA00] APA. "Publication Manual of the American Psychological Association". In: Sixth. American Psychological Association, 2000. Chap. General Guidelines for Reducing Bias.

[Ara+12] Pablo Aragón *et al.* "Biographical social networks on Wikipedia: a cross-cultural study of links that made history". In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM. 2012, p. 19.

[Arc+11] Daniel Archambault *et al.* "ThemeCrowds: Multiresolution summaries of twitter usage". In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM. 2011, pp. 77–84.

[Asc46] Solomon E Asch. "Forming impressions of personality." In: *The Journal of Abnormal and Social Psychology* 41.3 (1946), p. 258.

[ASL12] Maik Anderka, Benno Stein, and Nedim Lipka. "Predicting quality flaws in user-generated content: the case of Wikipedia". In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 981–990.

[BA99] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *Science* 286.5439 (1999), pp. 509–512.

[Bae05] Ricardo Baeza-Yates. "Applications of web query mining". In: *Advances in Information Retrieval*. Springer, 2005, pp. 7–22.

[Bar+12] Matías Barahona *et al.* "Tracking the 2011 student-led movement in chile through social media use". In: *Collective Intelligence 2012* (2012).

[Bar10] Shaowen Bardzell. "Feminist HCI: taking stock and outlining an agenda for design". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1301–1310.

[Bar15] Pablo Barberá. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political Analysis* 23.1 (2015), pp. 76–91.

[BCE07] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis N Efthimiadis. "Characterization of national web domains". In: *ACM Transactions on Internet Technology (TOIT)* 7.2 (2007), p. 9.

[Ben+09] Fabrício Benevenuto *et al.* "Characterizing user behavior in online social networks". In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM. 2009, pp. 49–62.

[BF05] Ulrik Brandes and Daniel Fleischer. *Centrality measures based on current flow*. Springer, 2005.

[BHV00] Mark Bruls, Kees Huizing, and Jarke J Van Wijk. *Squarified treemaps*. Springer, 2000.

[Big03] Brigitte Bigi. "Using Kullback-Leibler distance for text categorization". In: *Advances in Information Retrieval*. Springer, 2003, pp. 305–319.

[Bis+06] Christopher M Bishop *et al. Pattern recognition and machine learning*. Vol. 4. 4. Springer New York, 2006.

[BKN07] Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. *The adaptive web: methods and strategies of web personalization*. Vol. 4321. Springer Science & Business Media, 2007.

[BKY13] Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. "What's in Twitter, I know what parties are popular and who you are supporting now!" In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1379–1391.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[BOH11] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. "D^3 data-driven documents". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2301–2309.

[BP98] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine". In: *Computer networks and ISDN systems* 30.1 (1998), pp. 107–117.

[BR11a] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval: the concepts and technology behind search, 2nd. Edition*. Addison-Wesley, Pearson, 2011.

[BR11b] Michael J Brzozowski and Daniel M Romero. "Who Should I Follow? Recommending People in Directed Social Networks." In: *International Conference on Weblogs and Social Media*. 2011.

[Bra99] Mark Bray. "Control of education: Issues and tensions in centralization and decentralization". In: *Comparative education: The dialectic of the global and the local* (1999), pp. 207–232.

[BS03] Benjamin B Bederson and Ben Shneiderman. *The craft of information visualization: readings and reflections*. Morgan Kaufmann, 2003.

[BS14] David Bamman and Noah A Smith. "Unsupervised Discovery of Biographical Structure from Text". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 363–376.

[But06] Judith Butler. *Precarious life: The powers of mourning and violence.* Verso, 2006.

[Bux10] Bill Buxton. *Sketching User Experiences: Getting the Design Right and the Right Design.* Morgan Kaufmann, 2010.

[Cai12] Alberto Cairo. *The Functional Art: An introduction to information graphics and visualization.* New Riders, 2012.

[Can15] María Jesús Ibáñez Canelo. ""El control de los cuerpos de las mujeres es algo medular en la política patriarcal capitalista": entrevista a Soledad Rojas, feminista chilena". In: *Comunicación y Medios* 30 (2015).

[CB12] Benjamin Collier and Julia Bear. "Conflict, criticism, or confidence: an empirical examination of the gender gap in Wikipedia contributions". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.* ACM. 2012, pp. 383–392.

[CCL10] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users". In: *Proceedings of the 19th ACM international conference on Information and knowledge management.* ACM. 2010, pp. 759–768.

[CCM00] Chaomei Chen, Mary Czerwinski, and Robert Macredie. "Individual differences in virtual environments—introduction and overview". In: *Journal of the American Society for Information Science* 51.6 (2000), pp. 499–507.

[CEP13] CEP. *National Survey of Public Opinion, September–Octuber 2013.* http://www.cepchile.cl/1_5388/doc/estudio_nacional_de_opinion_publica_septiembre-octubre_2013.html. [Online; accessed April 2015]. 2013.

[CG04] Robert B Cialdini and Noah J Goldstein. "Social influence: Compliance and conformity". In: *Annu. Rev. Psychol.* 55 (2004), pp. 591–621.

[CG95] Kenneth W Church and William A Gale. "Poisson mixtures". In: *Natural Language Engineering* 1.02 (1995), pp. 163–190.

[CH11] Ewa S Callahan and Susan C Herring. "Cultural bias in Wikipedia content on famous persons". In: *Journal of the American society for information science and technology* 62.10 (2011), pp. 1899–1915.

[CH90] Kenneth Ward Church and Patrick Hanks. "Word association norms, mutual information, and lexicography". In: *Computational linguistics* 16.1 (1990), pp. 22–29.

[Cha+14] Shuo Chang *et al.* "Specialization, homophily, and gender in a social curation site: findings from pinterest". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 674–686.

[Che+09] Jilin Chen *et al.* "Make new friends, but keep the old: recommending people on social networking sites". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 201–210.

[Che+10] Jilin Chen *et al.* "Short and tweet: experiments on recommending content from information streams". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1185–1194.

[Che+12] Kailong Chen *et al.* "Collaborative personalized tweet recommendation". In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 661–670.

[CMS99] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. "Data preparation for mining world wide web browsing patterns". In: *Knowledge and information systems* 1.1 (1999), pp. 5–32.

[Coa04] Jennifer Coates. *Women, men, and language: A sociolinguistic account of gender differences in language*. Pearson Education, 2004.

[Con+11a] Michael D Conover *et al.* "Predicting the political alignment of twitter users". In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*. IEEE. 2011, pp. 192–199.

[Con+11b] Michael Conover *et al.* "Political polarization on twitter." In: *International Conference on Weblogs and Social Media*. 2011.

[CP11] Andreea S Calude and Mark Pagel. "How do we use language? Shared patterns in the frequency of word use across 17 world languages". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 366.1567 (2011), pp. 1101–1107.

[CR13a] Sidharth Chhabra and Paul Resnick. "Does clustered presentation lead readers to diverse selections?" In: *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2013, pp. 1689–1694.

[CR13b] Raviv Cohen and Derek Ruths. *Classifying Political Orientation on Twitter: It's Not Easy!* 2013.

[Cra+13] Andy Crabtree *et al.* "Introduction to the special issue of "The Turn to The Wild"". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 20.3 (2013), p. 13.

[CS03] Charles R Collins and Kenneth Stephenson. "A circle packing algorithm". In: *Computational Geometry* 25.3 (2003), pp. 233–256.

[CT14] Giovanni Luca Ciampaglia and Dario Taraborelli. "MoodBar: Increasing new user retention in Wikipedia through lightweight socialization". In: *arXiv preprint arXiv:1409.1496* (2014).

[Cum04] ML Cummings. "Automation bias in intelligent time critical decision support systems". In: *AIAA 1st Intelligent Systems Technical Conference*. Vol. 2. 2004, pp. 557–562.

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[DCC11] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. "Identifying relevant social media content: leveraging information diversity and user cognition". In: *Proceedings of the 22nd ACM conference on Hypertext and Hypermedia*. ACM. 2011, pp. 161–170.

[De 12] Simone De Beauvoir. *The second sex*. Random House LLC, 2012.

[DE07] Juan Carlos Dürsteler and Yuri Engelhardt. *The digital magazine of InfoVis.net, message n°187*. [Online; accessed April 2015]. InfoVis.net, 2007.

[DNK10] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. "Diamonds in the rough: Social media visual analytics for journalistic inquiry". In: *2010 IEEE Symposium on Visual Analytics Science and Technology (VAST)*. IEEE. 2010, pp. 115–122.

[Don+10] Judith Donath *et al.* "Data portraits". In: *ACM SIGGRAPH 2010 Art Gallery*. ACM. 2010, pp. 375–383.

[Don14] Judith Donath. *The social machine: designs for living online*. MIT Press, 2014.

[Dor+10] Marian Dork *et al.* "A visual backchannel for large-scale events". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1129–1138.

[Dra09] Alex Dragulescu. *Lexigraphs: Twitter Data Portrait: jakedfw*. http://vimeo.com/2404119. [Online; accessed April 2015]. 2009.

[Far+10] Siamak Faridani *et al.* "Opinion space: a scalable tool for browsing online comments". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1175–1184.

[Fes62] Leon Festinger. *A theory of Cognitive Dissonance*. Vol. 2. Stanford University Press, 1962.

[FFG14] Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. "What makes a good biography?: multidimensional quality analysis based on Wikipedia article feedback data". In: *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee. 2014, pp. 855–866.

[Fil13] Amanda Filipacchi. "Wikipedia's sexism toward female novelists". In: *The New York Times, April 28th, 2013* (2013).

[FM12] Michela Ferron and Paolo Massa. "Psychological processes underlying Wikipedia representations of natural and manmade disasters". In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM. 2012, p. 2.

[FN90] Susan T Fiske and Steven L Neuberg. "A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation". In: *Advances in experimental social psychology* 23 (1990), pp. 1–74.

[For+07] S. Fortunato *et al.* "On local estimations of PageRank: A mean field approach". In: *Internet Mathematics* 4.2–3 (2007), pp. 245–266. DOI: 10.1080/15427951.2007.10129294.

[FPS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases". In: *AI magazine* 17.3 (1996), p. 37.

[Fre04] Linton Freeman. "The development of social network analysis". In: *A Study in the Sociology of Science* (2004).

[Fre77] Linton C Freeman. "A set of measures of centrality based on betweenness". In: *Sociometry* (1977), pp. 35–41.

[Fri02] Michael Friendly. "Visions and re-visions of Charles Joseph Minard". In: *Journal of Educational and Behavioral Statistics* 27.1 (2002), pp. 31–51.

[Fri10] Betty Friedan. *The feminine mystique*. WW Norton & Company, 2010.

[Fry00] Benjamin Jotham Fry. "Organic Information Design". PhD thesis. Massachusetts Institute of Technology, 2000.

[GH11] Jennifer Golbeck and Derek Hansen. "Computing political preference among twitter followers". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 1105–1108.

[Gid+00] Anthony Giddens *et al. Introduction to sociology*. WW Norton New York, 2000.

[Gil05] Jim Giles. "Internet encyclopaedias go head to head". In: *Nature* 438.7070 (2005), pp. 900–901.

[GK08] Sebastian Galiani and Sukkoo Kim. "Political Centralization and Urban Primacy: Evidence from National and Provincial Capitals in the Americas". In: *Understanding Long-Run Economic Growth: Geography, Institutions, and the Knowledge Economy*. University of Chicago Press, 2008, pp. 121–153.

[GMQ14] Ruth Garcia-Gavilanes, Yelena Mejova, and Daniele Quercia. "Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 1511–1522.

[Goe+13] Ashish Goel *et al.* "Discovering Similar Users on Twitter". In: *11th Workshop on Mining and Learning with Graphs*. 2013.

[Gof59] Erving Goffman. "The presentation of self in everyday life". In: (1959).

[Gon+14] Sandra González-Bailón *et al.* "Assessing the bias in samples of large online networks". In: *Social Networks* 38 (2014), pp. 16–27.

[Gou+11] Liang Gou *et al.* "Sfviz: interest-based friends exploration and recommendation in social networks". In: *Proceedings of the 2011 Visual Information Communication-International Symposium*. ACM. 2011, p. 15.

[GR89] Andrew Gillespie and Kevin Robins. "Geographical inequalities: The spatial bias of the new communications technologies". In: *Journal of Communication* 39.3 (1989), pp. 7–18.

[Gre+10] Brynjar Gretarsson *et al.* "Smallworlds: Visualizing social recommendations". In: *Computer Graphics Forum*. Vol. 29. 3. Wiley Online Library. 2010, pp. 833–842.

[Gre+12] Brynjar Gretarsson *et al.* "Topicnets: Visual analysis of large text corpora with topic modeling". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.2 (2012), p. 23.

[Gup+13] Pankaj Gupta *et al.* "Wtf: The who to follow service at twitter". In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 505–514.

[Har+09] William Hart *et al.* "Feeling validated versus being correct: a meta-analysis of selective exposure to information." In: *Psychological bulletin* 135.4 (2009), p. 555.

[HB03] Mark Harrower and Cynthia A Brewer. "ColorBrewer. org: an online tool for selecting colour schemes for maps". In: *The Cartographic Journal* 40.1 (2003), pp. 27–37.

[Hb05] Jeffrey Heer and danah boyd. "Vizster: Visualizing Online Social Networks". In: *IEEE Information Visualization (InfoVis)*. 2005, pp. 32–39. URL: http://vis.stanford.edu/papers/vizster.

[HBO10] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. "A tour through the visualization zoo." In: *Commun. ACM* 53.6 (2010), pp. 59–67.

[HBS10] John Hannon, Mike Bennett, and Barry Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches". In: *Proceedings of the fourth ACM Conference on Recommender Systems*. ACM. 2010, pp. 199–206.

[Hec+11] Brent Hecht *et al.* "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles". In: *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM. 2011, pp. 237–246.

[Her+04] Jonathan L Herlocker *et al.* "Evaluating collaborative filtering recommender systems". In: *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), pp. 5–53.

[HG09] Brent Hecht and Darren Gergle. "Measuring self-focus bias in community-maintained knowledge repositories". In: *Proceedings of the fourth international conference on Communities and technologies*. ACM. 2009, pp. 11–20.

[HHM10] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and organizations, software of the mind: intercultural cooperation and its importance for survival (3rd edition)*. McGraw-Hill Professional, 2010.

[HKR00] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. "Explaining collaborative filtering recommendations". In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM. 2000, pp. 241–250.

[HMS13] Itai Himelboim, Stephen McCreery, and Marc Smith. "Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter". In: *Journal of Computer-Mediated Communication* 18.2 (2013), pp. 40–60.

[HN98] Jerry L Hintze and Ray D Nelson. "Violin plots: a box plot-density trace synergism". In: *The American Statistician* 52.2 (1998), pp. 181–184.

[HNA05] Martie G Haselton, Daniel Nettle, and Paul W Andrews. "The evolution of cognitive bias". In: *The handbook of evolutionary psychology* (2005), pp. 724–746.

[HS13] Benjamin Mako Hill and Aaron Shaw. "The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation". In: *PloS ONE* 8.6 (2013), e65782.

[HYG13] CJ Hutto, Sarita Yardi, and Eric Gilbert. "A longitudinal study of follow predictors on twitter". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, pp. 821–830.

[Ios+14] Daniela Iosub *et al.* "Emotions under discussion: Gender, status and communication in online collaboration". In: *PloS ONE* 9.8 (2014), e104880.

[Jac89] Jerry A Jacobs. *Revolving doors: Sex segregation and women's careers*. Stanford University Press, 1989.

[Jos06] Lou Jost. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[JS91] Brian Johnson and Ben Shneiderman. "Tree-maps: A space-filling approach to the visualization of hierarchical information structures". In: *IEEE Conference on Visualization*. IEEE. 1991, pp. 284–291.

[Kei+08] Daniel Keim *et al. Visual analytics: Definition, process, and challenges*. Springer, 2008.

[KFS06] Shipra Kayan, Susan R Fussell, and Leslie D Setlock. "Cultural differences in the use of instant messaging in Asia and North America". In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM. 2006, pp. 525–528.

[Kle99] Jon M Kleinberg. "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[KMO11] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. "I'm eating a sandwich in Glasgow: modeling locations with tweets". In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM. 2011, pp. 61–68.

[KNM10] Maurits Clemens Kaptein, Clifford Nass, and Panos Markopoulos. "Powerful and consistent analysis of likert-type ratingscales". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 2391–2394.

[Kol13] Ken Kollman. *Perils of centralization: lessons from church, state, and corporation*. Cambridge University Press, 2013.

[Kon10] Piotr Konieczny. "Teaching with Wikipedia and other Wikimedia foundation wikis". In: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*. ACM. 2010, p. 29.

[Kri+12] Travis Kriplean *et al.* "Supporting reflective public thought with ConsiderIt". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM. 2012, pp. 265–274.

[Kru99] Paul Krugman. "The role of geography in development". In: *International regional science review* 22.2 (1999), pp. 142–161.

[KRW11] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. "Each to his own: how different users call for different interaction methods in recommender systems". In: *Proceedings of the fifth ACM conference on Recommender systems*. ACM. 2011, pp. 141–148.

[Kul+12] Juhi Kulshrestha *et al.* "Geographic dissection of the Twitter network". In: *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. 2012.

[Kwa+10] Haewoon Kwak *et al.* "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 591–600.

[L+54] Paul F Lazarsfeld, Robert K Merton, *et al.* "Friendship as a social process: A substantive and methodological analysis". In: *Freedom and control in modern society* 18.1 (1954), pp. 18–66.

[Lak73] Robin Tolmach Lakoff. "Language and woman's place". In: *Language in Society* 2, No. 1, Apr. (1973), pp. 45–80.

[Lam+11] Shyong Tony K Lam *et al.* "WP: clubhouse?: an exploration of Wikipedia's gender imbalance". In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM. 2011, pp. 1–10.

[Lam+12] Heidi Lam *et al.* "Empirical studies in information visualization: Seven scenarios". In: *IEEE Transactions on Visualization and Computer Graphics* 18.9 (2012), pp. 1520–1536.

[Lan+12] David Laniado *et al.* "Emotions and dialogue in a peer-production community: the case of Wikipedia". In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM. 2012, p. 9.

[LB12] Marco Lui and Timothy Baldwin. "langid.py: An off-the-shelf language identification tool". In: *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics. 2012, pp. 25–30.

[Leh+14a] Janette Lehmann *et al.* "Reader preferences and behavior on Wikipedia". In: *Proceedings of the 25th ACM conference on Hypertext and Social Media*. ACM. 2014, pp. 88–97.

[Leh+14b] Jens Lehmann *et al.* "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". In: *Semantic Web Journal* (2014).

[LF13] Q Vera Liao and Wai-Tat Fu. "Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information". In: *Proceedings of the ACM CHI*. 2013, pp. 2359–2368.

[LF14] Q Vera Liao and Wai-Tat Fu. "Can you hear me now?: mitigating the echo chamber effect by source position indicators". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 184–196.

[Lim12] Merlyna Lim. "Clicks, cabs, and coffee houses: Social media and oppositional movements in Egypt, 2004–2011". In: *Journal of Communication* 62.2 (2012), pp. 231–248.

[Liu+14] Shixia Liu *et al.* "A survey on information visualization: recent advances and challenges". In: *The Visual Computer* 30.12 (2014), pp. 1373–1393.

[LOY14] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. *Measuring User Engagement*. Morgan & Claypool Publishers, 2014.

[LS10] Zhicheng Liu and John T Stasko. "Mental models, visual reasoning and interaction in information visualization: A top-down perspective". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 999–1008.

[Luc+13] Andrés Lucero *et al.* "The playful experiences (plex) framework as a guide for expert evaluation". In: *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces*. ACM. 2013, pp. 221–230.

[Lug08] Jairo Lugo. *The Media in Latin America*. McGraw-Hill International, 2008.

[Lyn72] Kevin Lynch. *What Time is This Place?* The MIT Press, 1972.

[Mac+11] Alan M MacEachren *et al.* "Senseplace2: Geotwitter analytics support for situational awareness". In: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE. 2011, pp. 181–190.

[Mar+11] Adam Marcus *et al.* "Twitinfo: aggregating and visualizing microblogs for event exploration". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 227–236.

[Mar03] Noah P Mark. "Culture and competition: Homophily and distancing explanations for cultural niches". In: *American Sociological Review* (2003), pp. 319–345.

[McC80] Peter McCullagh. "Regression models for ordinal data". In: *Journal of the royal statistical society. Series B (Methodological)* (1980), pp. 109–142.

[Mei13] Isabel Meirelles. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport publishers, 2013.

[Mik+13] Tomas Mikolov *et al.* "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119.

[Mil67] Stanley Milgram. "The small world problem". In: *Psychology today* 2.1 (1967), pp. 60–67.

[Min11] Gobierno de Chile Ministerio de Desarrollo Social. *CASEN Survey*. http://observatorio.ministeriodesarrollosocial.gob.cl/casen_obj.php. [In spanish; Online; accessed April 2015]. 2011.

[Mis+07] Alan Mislove *et al.* "Measurement and analysis of online social networks". In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM. 2007, pp. 29–42.

[Mis+11] Alan Mislove *et al.* "Understanding the demographics of Twitter users". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain*. 2011.

[ML75] David G Myers and Helmut Lamm. "The polarizing effect of group discussion: The discovery that discussion tends to enhance the average prediscussion tendency has stimulated new insights about the nature of group influence". In: *American Scientist* (1975), pp. 297–303.

[MLR13] Sean A Munson, Stephanie Y Lee, and Paul Resnick. "Encouraging Reading of Diverse Political Viewpoints with a Browser Widget". In: *International AAAI Conference on Weblogs and Social Media* (2013).

[MM10] Matthew Michelson and Sofus A Macskassy. "Discovering users' topics of interest on twitter: a first look". In: *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. ACM. 2010, pp. 73–80.

[Mor+13] Jonathan T Morgan *et al.* "Tea and sympathy: crafting positive new user experiences on Wikipedia". In: *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM. 2013, pp. 839–848.

[MPC10] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. "Twitter under crisis: Can we trust what we RT?" In: *Proceedings of the first workshop on social media analytics*. ACM. 2010, pp. 71–79.

[MR10] Sean A Munson and Paul Resnick. "Presenting diverse political opinions: how and how much". In: *Proceedings of the ACM CHI*. 2010, pp. 1457–1466.

[MRK06] Sean M McNee, John Riedl, and Joseph A Konstan. "Being accurate is not enough: how accuracy metrics have hurt recommender systems". In: *CHI'06 extended abstracts on Human factors in computing systems*. ACM. 2006, pp. 1097–1101.

[MS96] Robert K Merton and Piotr Sztompka. *On social structure and science*. University of Chicago Press, 1996.

[MSC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* (2001), pp. 415–444.

[Mun09] Tamara Munzner. "A nested model for visualization design and validation". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 921–928.

[Mun12] Sean A Munson. "Exposure to political diversity online". PhD thesis. The University of Michigan, 2012.

[MZR09] Sean A Munson, Daniel Xiaodan Zhou, and Paul Resnick. "Sidelines: An Algorithm for Increasing Diversity in News and Opinion Aggregators". In: *International Conference on Weblogs and Social Media*. 2009.

[Nat14] National Statistics Office. *Country and Regional Populations of Chile: Updated 2002–2012, Projected 2013–2020*. http://www.ine.cl/canales/sala_prensa/noticias/noticia.php?opc=news&id=615&lang=esp. [Online; in Spanish; accessed April 2015]. 2014.

[NBG02] Brian A Nosek, Mahzarin Banaji, and Anthony G Greenwald. "Harvesting implicit group attitudes and beliefs from a demonstration web site." In: *Group Dynamics: Theory, Research, and Practice* 6.1 (2002), p. 101.

[NBL10] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. "Is it really about me?: message content in social awareness streams". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM. 2010, pp. 189–192.

[New05] Mark EJ Newman. "A measure of betweenness centrality based on random walks". In: *Social networks* 27.1 (2005), pp. 39–54.

[Nic98] Raymond S Nickerson. "Confirmation bias: A ubiquitous phenomenon in many guises." In: *Review of general psychology* 2.2 (1998), p. 175.

[Nus95] Martha C Nussbaum. "Objectification". In: *Philosophy & Public Affairs* 24.4 (1995), pp. 249–291.

[Oko+14] Chitu Okoli *et al.* "Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership". In: *Journal of the American Society for Information Science and Technology* (2014).

[Pag+99] Lawrence Page *et al. The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[Par+09] Souneil Park *et al.* "NewsCube: delivering multiple aspects of news to mitigate media bias". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM. 2009, pp. 443–452.

[Par11] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* Penguin UK, 2011.

[Pas10] Victor Pascual Cid. "Visual Exploration of Web Spaces". PhD thesis. Universitat Pompeu Fabra, 2010.

[PB15] Denis Parra and Peter Brusilovsky. "User-controllable personalization: A case study with SetFusion". In: *International Journal of Human-Computer Studies* (2015).

[PB79] David G Perry and Kay Bussey. "The social learning theory of sex differences: Imitation is alive and well." In: *Journal of Personality and Social Psychology* 37.10 (1979), p. 1699.

[Ped+11] F. Pedregosa *et al.* "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[PFB01] James W Pennebaker, Martha E Francis, and Roger J Booth. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71 (2001), p. 2001.

[PHK] Felicia Pratto, Peter J Hegarty, and Josephine D Korchmaros. "How communication practices and category norms lead people to stereotype particular people and groups". In: *Stereotype dynamics: Language based approaches to the formation, maintenance, and transformation of stereotypes* (), pp. 293–313.

[Pia14] Steven T Piantadosi. "Zipf's word frequency law in natural language: A critical review and future directions". In: *Psychonomic bulletin & review* (2014), pp. 1–19.

[Pob+11] Bárbara Poblete *et al.* "Do all birds tweet the same?: characterizing Twitter around the world". In: *Proceedings of the 20th ACM international conference on Information and knowledge management.* ACM. 2011, pp. 1025–1030.

[Pon14] Pontificia Universidad Católica de Chile. *Encuesta Bicentenario UC-Adimark, 2014*. http://encuestabicentenario.uc.cl/. [In Spanish; Online; accessed April 2015]. 2014.

[PP11] Marco Pennacchiotti and Ana-Maria Popescu. "Democrats, republicans and starbucks afficionados: user classification in twitter". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM. 2011, pp. 430–438.

[PSM07] Zachary Pousman, John T Stasko, and Michael Mateas. "Casual information visualization: Depictions of data in everyday life". In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (2007), pp. 1145–1152.

[QAC12] Daniele Quercia, Harry Askham, and Jon Crowcroft. "TweetLDA: supervised topic classification and link prediction in Twitter". In: *Proceedings of the 4th Annual ACM Web Science Conference*. ACM. 2012, pp. 247–250.

[QCC12] Daniele Quercia, Licia Capra, and Jon Crowcroft. "The social world of Twitter: Topics, geography, and emotions". In: *The 6th international AAAI Conference on weblogs and social media*. 2012.

[QJ80] George A Quattrone and Edward E Jones. "The perception of variability within in-groups and out-groups: Implications for the law of small numbers." In: *Journal of Personality and Social Psychology* 38.1 (1980), p. 141.

[Rat+10] Jacob Ratkiewicz *et al.* "Characterizing and modeling the dynamics of online popularity". In: *Physical review letters* 105.15 (2010), p. 158701.

[RB12] Luz Rello and Ricardo A Baeza-Yates. "Social Media Is NOT that Bad! The Lexical Quality of Social Media." In: *International Conference on Weblogs and Social Media*. 2012.

[RDL10] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. "Characterizing Microblogs with Topic Models". In: *International Conference on Weblogs and Social Media*. 2010.

[Rel14] Luz Rello. "DysWebxia: a text accessibility model for people with dyslexia". PhD thesis. Universitat Pompeu Fabra, 2014.

[RK04] Ryan Rifkin and Aldebaro Klautau. "In defense of one-vs-all classification". In: *The Journal of Machine Learning Research* 5 (2004), pp. 101–141.

[Ros06] Roy Rosenzweig. "Can history be open source? Wikipedia and the future of the past". In: *The Journal of American History* 93.1 (2006), pp. 117–146.

[Rou+13] Dominic Rout *et al.* "Where's@ wally?: a classification approach to geolocating users based on their social ties". In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM. 2013, pp. 11–20.

[RR11] Joseph Reagle and Lauren Rhue. "Gender bias in Wikipedia and Britannica". In: *International Journal of Communication* 5 (2011), p. 21.

[ŘS10] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 22, 2010, pp. 45–50.

[Sav+14] Saiph Savage *et al.* "Visualizing targeted audiences". In: *COOP 2014-Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France)*. Springer. 2014, pp. 17–34.

[SB07] Bonnie L Shepard and Lidia Casas Becerra. "Abortion policies and practices in Chile: ambiguities and dilemmas". In: *Reproductive Health Matters* 15.30 (2007), pp. 202–210.

[Sch+10] Emanuel Schmider *et al.* "Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption." In: *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6.4 (2010), p. 147.

[Seo+13] Hyunjeong Seo *et al.* "Network-based approaches for anticancer therapy (Review)". In: *International Journal of Oncology* 43.6 (2013), pp. 1737–1744.

[SFM09] M Ángeles Serrano, Alessandro Flammini, and Filippo Menczer. "Modeling statistical properties of written text". In: *PloS ONE* 4.4 (2009), e5372.

[SH91] William Strauss and Neil Howe. *Generations: The history of America's future, 1584 to 2069*. Morrow New York, NY: 1991.

[Shn96] Ben Shneiderman. "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE. 1996, pp. 336–343.

[SM93] Lynn Smith-Lovin and J Miller McPherson. "You are who you know: A network approach to gender". In: *Theory on gender/feminism on theory* (1993), pp. 223–51.

[Sri+00] Jaideep Srivastava *et al.* "Web usage mining: Discovery and applications of usage patterns from web data". In: *ACM SIGKDD Explorations Newsletter* 1.2 (2000), pp. 12–23.

[Sti13] Sarah Stierch. *Women and Wikimedia Survey 2011*. https://meta.wikimedia.org/wiki/Women_and_Wikimedia_Survey_2011. [Online; accessed April 2015]. 2013.

[Sun09] Cass R Sunstein. *Going to extremes: How like minds unite and divide*. Oxford University Press, 2009.

[SW14] Steven S. Skiena and Charles B. Ward. *Who's Bigger?: Where Historical Figures Really Rank*. Cambridge Univ. Press, 2014.

[SWW07] Toni Schmader, Jessica Whitehead, and Vicki H Wysocki. "A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants". In: *Sex Roles* 57.7-8 (2007), pp. 509–514.

[SZ89] Karen Stephenson and Marvin Zelen. "Rethinking centrality: Methods and examples". In: *Social Networks* 11.1 (1989), pp. 1–37.

[Tan01] Junichirō Tanizaki. *In praise of shadows*. Random House, 2001.

[TGW12] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. "Geography of Twitter networks". In: *Social networks* 34.1 (2012), pp. 73–81.

[TPL06] Matt Thomas, Bo Pang, and Lillian Lee. "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts". In: *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2006, pp. 327–335.

[Uni] United Nations. *Abortion Policies. A Global Review.* http://www.un.org/esa/population/publications/abortion/profiles.htm. [Online; accessed April 2015].

[Van06] Jarke J Van Wijk. "Bridging the gaps". In: *Computer Graphics and Applications* 26.6 (2006), pp. 6–9.

[VGD06] Fernanda B Viégas, Scott Golder, and Judith Donath. "Visualizing email content: portraying relationships from conversational histories". In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pp. 979–988.

[Vié+13] Fernanda Viégas *et al.* "Google+ Ripples: a native visualization of information flow". In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 1389–1398.

[Vog79] Helmut Vogel. "A better way to construct the sunflower head". In: *Mathematical biosciences* 44.3 (1979), pp. 179–189.

[VW08] Fernanda B Viégas and Martin Wattenberg. "Tag clouds and the case for vernacular visualization". In: *Interactions* 15.4 (2008), pp. 49–52.

[VWF09] Fernanda B Viegas, Martin Wattenberg, and Jonathan Feinberg. "Participatory visualization with wordle". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1137–1144.

[VWV09] Frank Van Ham, Martin Wattenberg, and Fernanda B Viégas. "Mapping text with Phrase Nets". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1169–1176.

[Wag+15] Claudia Wagner *et al.* "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia". In: *arXiv preprint arXiv:1501.06307* (2015).

[Wat13] Duncan J Watts. "Computational social science: Exciting progress and future directions". In: *The Bridge on Frontiers of Engineering* 43.4 (2013), pp. 5–10.

[Wen+10] Jianshu Weng *et al.* "Twitterrank: finding topic-sensitive influential twitterers". In: *Proceedings of the third ACM WSDM*. 2010, pp. 261–270.

[Wes04] Marcos Weskamp. *newsmap*. http://newsmap.jp. [Online; accessed 21-March-2014]. 2004.

[WGB13] Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. "Secular vs. Islamist Polarization in Egypt on Twitter". In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM. 2013, pp. 290–297.

[Wik15] Wikipedia. *Abortion debate*. https : / / en . wikipedia . org / wiki / Abortion_debate. [Online; accessed April 2015]. 2015.

[Wil+12] Christo Wilson *et al.* "Beyond social graphs: User interactions in online social networks and their implications". In: *ACM Transactions on the Web (TWEB)* 6.4 (2012), p. 17.

[WK06] Martin Wattenberg and Jesse Kriss. "Designing for social data analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 12.4 (2006), pp. 549–557.

[WS14] Charlotte Witt and Lisa Shapiro. "Feminist History of Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2014. 2014.

[WS98] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (1998), pp. 440–442.

[WV08] Martin Wattenberg and Fernanda B Viégas. "The Word Tree, an interactive visual concordance". In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1221–1228.

[XD99] Rebecca Xiong and Judith Donath. "PeopleGarden: creating data portraits for users". In: *Proceedings of the 12th annual ACM symposium on User interface software and technology*. ACM. 1999, pp. 37–44.

[YB10] Sarita Yardi and Danah Boyd. "Dynamic debates: An analysis of group polarization over time on twitter". In: *Bulletin of Science, Technology & Society* 30.5 (2010), pp. 316–327.

[YDG13] Elad Yom-Tov, Susan Dumais, and Qi Guo. "Promoting civil discourse through search engine diversity". In: *Social Science Computer Review* (2013), p. 0894439313506838.

[Yi+08] Ji Soo Yi *et al.* "Understanding and characterizing insights: how do people gain insights using information visualization?" In: *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*. ACM. 2008, p. 4.

[Yom+12] Elad Yom-Tov *et al.* "Pro-anorexia and pro-recovery photo sharing: a tale of two warring tribes". In: *Journal of Medical Internet Research* 14.6 (2012).

[Zie+05] Cai-Nicolas Ziegler *et al.* "Improving recommendation lists through topic diversification". In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 22–32.

[Zip49] George Kingsley Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[Zla+06] Vinko Zlatić *et al.* "Wikipedias: Collaborative web-based encyclopedias as complex networks". In: *Physical Review E* 74.1 (2006), p. 016115.